REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
28-06-2016	Final Report	17-Sep-2012 - 16-Mar-2016

4. TITLE AND SUBTITLE
Final Report: Detection and Interpretation of Low-Level and High-Level Surprising and Important Events in Large-Scale Data Streams

5a. CONTRACT NUMBER
W911NF-12-1-0433

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Laurent Itti

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Southern California
Department of Contracts and Grants
3720 South Flower Street
Los Angeles, CA 90089 -0701

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
62221-NS.13

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
This project explored how to mathematically formalize the computations of surprise and relevance of events in large data streams, including video, audio and text. We have developed new mathematical theories to define surprise in terms of how new data observations may or not affect an observer's set of beliefs. This is computed in terms of the Kullback-Leibler divergence between posterior and prior beliefs of the observer, and quantified in a new unit of "wows". Likewise, we have developed a new general theory of relevance that quantifies how new data observations may or not affect an observer's beliefs about how she/he/it will achieve its goals. Data observations

15. SUBJECT TERMS
surprise; attention; social media; web search; relevance; machine vision

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Laurent Itti
UU	UU	UU	UU		19b. TELEPHONE NUMBER 213-740-3527

Report Title

Final Report: Detection and Interpretation of Low-Level and High-Level Surprising and Important Events in Large-Scale Data Streams

ABSTRACT

This project explored how to mathematically formalize the computations of surprise and relevance of events in large data streams, including video, audio and text. We have developed new mathematical theories to define surprise in terms of how new data observations may or not affect an observer's set of beliefs. This is computed in terms of the Kullback-Leibler divergence between posterior and prior beliefs of the observer, and quantified in a new unit of "wows". Likewise, we have developed a new general theory of relevance that quantifies how new data observations may or not affect an observer's beliefs about how she/he/it will achieve its goals. Data observations which suggest that some previously possible solutions to a problem are now invalid will be measured as more relevant, in a new unit of "rels". Both theories have been extensively tested using large video (~3000 hours) and text (twitter feeds) datasets.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received		Paper
08/30/2013	2.00	Ali Borji, Dicky N. Sihite, Laurent Itti. What stands out in a scene? A study of human explicit saliency judgment, Vision Research, (10 2013): 0. doi: 10.1016/j.visres.2013.07.016
08/30/2013	3.00	A Borji, D N Sihite, L Itti. Objects do not predict fixations better than early saliency: A re-analysis of Einhaeuser et al.'s data, Journal of Vision, (07 2013): 1. doi:
09/02/2014	5.00	A. Borji, L. Itti. Defending Yarbus: Eye movements reveal observers' task, Journal of Vision, (03 2014): 1. doi: 10.1167/14.3.29
09/02/2014	6.00	Christian Siagian, Chin Kai Chang, Laurent Itti. Autonomous Mobile Robot Localization and Navigation Using a Hierarchical Map Representation Primarily Guided by Vision, Journal of Field Robotics, (05 2014): 408. doi: 10.1002/rob.21505
10/08/2015	10.00	Daniel Parks, Ali Borji, Laurent Itti. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes, Vision Research, (11 2014): 0. doi: 10.1016/j.visres.2014.10.027
10/08/2015	11.00	A. Borji, D. Parks, L. Itti. Complementary effects of gaze direction and early saliency in guiding fixations during free viewing, Journal of Vision, (11 2014): 0. doi: 10.1167/14.13.3
TOTAL:		**6**

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received Paper

TOTAL:

Number of Manuscripts:

Books

Received Book

TOTAL:

Received Book Chapter

08/30/2013	4.00	L Itti, A Borji. Computational models: Bottom-up and top-down aspects, Oxford, UK: Oxford University Press, (06 2013)
08/31/2013	1.00	A Borji, L Itti. Computational models of attention, New York, NY: W. W. Norton & Company, (01 2014)
10/08/2015	9.00	Farhan Baluch, Laurent Itti. Mining Videos for Features that Drive Attention, Springer International Publishing: Springer International Publishing, (05 2015)

TOTAL: **3**

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Chen Zhang	0.30	
Randolph Voorhies	0.30	
Daniel F Parks	0.30	
FTE Equivalent:	**0.90**	
Total Number:	**3**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: 0.00

Names of Personnel receiving masters degrees

NAME
Total Number:

Names of personnel receiving PHDs

NAME	
Randolph Voorhies	
Daniel F Parks	
Total Number:	**2**

Names of other research staff

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

See Attachment

Technology Transfer

Final progress report

"Detection and Interpretation of Low-Level and High-Level Surprising and Important Events in Large-Scale Data Streams"

PI: Prof. Laurent Itti, USC

ARO proposal number: 62221-NS, award number: W911NF-12-1-0433

Summary: This project explored how to mathematically formalize the computations of surprise and relevance of events in large data streams, including video, audio and text. We have developed new mathematical theories to define surprise in terms of how new data observations may or not affect an observer's set of beliefs. This is computed in terms of the Kullback-Leibler divergence between posterior and prior beliefs of the observer, and quantified in a new unit of "wows". Likewise, we have developed a new general theory of relevance that quantifies how new data observations may or not affect an observer's beliefs about how she/he/it will achieve its goals. Data observations which suggest that some previously possible solutions to a problem are now invalid will be measured as more relevant, in a new unit of "rels". Both theories have been extensively tested using large video (~3000 hours) and text (twitter feeds) datasets.

Publications:

J. Zhao, C. Siagian, L. Itti, Fixation Bank: Learning to Reweight Fixation Candidates, In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp. 3174-3182, Jun 2015. [2015 acceptance rate: 28%]

F. Baluch, L. Itti, Mining videos for features that drive attention, In: Multimedia Data Mining and Analytics, (A. K. Baughman, J. Gao, J. Pan, V. A. Petrushin Ed.), pp. 311-326, Apr 2015.

D. Parks, A. Borji, L. Itti, Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes, Vision Research, pp. 1-12, 2015. [2013 impact factor: 2.381]

A. Borji, D. Parks, L. Itti, Complementary effects of gaze direction and early saliency in guiding fixations during free viewing, Journal of Vision, Vol. 14, No. 13, pp. 1-32, Nov 2014. [2013 impact factor: 2.727]

A. Borji, L. Itti, Defending Yarbus: Eye movements reveal observers' task, Journal of Vision, Vol. 14, No. 3(29), pp. 1-22, Mar 2014. [2012 impact factor: 2.47]

C. Siagian, C.-K. Chang, L. Itti, Autonomous Mobile Robot Localization and Navigation Using Hierarchical Map Representation Primarily Guided by Vision, Journal of Field Robotics, Vol. 31, No. 3, pp. 408-440, May/Jun 2014. [2012 impact factor: 2.152]

J. Windau, L. Itti, Situation awareness via sensor-equipped eyeglasses, In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5674-5679, Nov 2013. [2013 acceptance rate: 43%]

A. McNamara, K. Mania, G. Koulieris, L. Itti, Attention-aware rendering, mobile graphics and games, In: Proceeding ACM SIGGRAPH 2014 Courses, pp. 6.1 - 6.119, Aug 2014.

L. Itti, A. Borji, Computational models of attention, In: Cognitive Neuroscience: The Biology of the Mind (Fifth Edition), (M. S. Gazzaniga, R. B. Ivry, G. R. Mangun Ed.), pp. 1-10, 2014.

A. Borji, D. N. Sihite, L. Itti, Objects do not predict fixations better than early saliency: A re-analysis of Einhaeuser et al.'s data, Journal of Vision, Vol. 13, No. 10, pp. 1-4, Aug 2013. [2011 impact factor: 2.47]

A. Borji, D. N. Sihite, L. Itti, What stands out in a scene? A study of human explicit saliency judgment, Vision Research, Vol. 91, pp. 62-77, Aug 2013. [2011 impact factor: 2.13]

L. Itti, A. Borji, Computational models: Bottom-up and top-down aspects, In: The Oxford Handbook of Attention, (A. C. Nobre, S. Kastner Ed.), pp. 1-20, 2013.

The main goals of this project are to investigate the concepts of ***surprise*** and ***relevance*** in data streams that include video, audio, and text. In year 1, we made significant progress on surprise in video, demonstrating breakthrough results with finding surprising events in a large corpus of surveillance video streams (cameras on the Texas/Mexico border). In year 2, we introduced a new mathematical definition of goal relevance that is built from first principles, general and widely applicable. In year 3, we have focused on computing surprise on completely different data streams (web pages and social media feeds), and on applications to visual attention modeling.

1. Defining and modeling surprise

In year 1, we focused on two major thrusts: detecting surprising events in video streams from surveillance cameras, and building one of the first computational models of task-driven guidance of visual attention and gaze towards items that are more relevant to one's goals. We have achieved breakthrough results on both fronts. With respect to surprising events in video, we have created a system that can monitor outdoor video streams and detect surprising events, then summarizing the videos into short segments around each event. We obtained 2,783 videos from 24 different cameras (color or infrared) placed along the Texas-Mexico border. While the full dataset is being ground-truthed, we processed a preliminary subset of 89 videos from 13 cameras, totaling 4.8M frames or 89 hours. Our system correctly dismissed over 99% of all frames as being boring, while detecting 56 surprising events (later manually classified as 16 *interesting/relevant* events such as cars, boats, etc and 40 birds), with only 4 false alarms (due to camera glitches) and 2 misses (1 truck and 1 boat). With respect to top-down attention towards items that are relevant to one's task, we developed a new Bayesian framework that combines information from past gaze fixation points, past video frames, and past actions to guess where a human observer may look next given a task. The system was tested with 3D interactive video games including driving, flight combat, or running a hotdog stand that serves many hungry customers. It was able to predict where the human players would look next, significantly above chance and better than the state of the art. This is the first neuro-computational model of attention that we know of which reasons about objects, scene context, past gaze points and past actions to determine the next most task-relevant location to look at.

Several types of feature detectors were chosen for our implementation, all of which have been well documented in the literature on the primary visual cortex of mammals. Each detector type is implemented as approximations to center-surround cells with receptive fields spanning a range of sizes. This approximation is efficiently performed by constructing Gaussian image pyramids spanning from 2 octaves to 8 octaves above the original scale, for a variety of feature types. In total, we create 12 different low-level feature pyramids (one intensity, two color, four orientation, four motion, one flicker), which are combined to create a total of 72 center surround feature detector maps.

Surprise computation is applied to each of the 72 maps. A schematic diagram of a single surprise computation unit is shown in fig. 1, which takes as input a prior distribution and a data distribution and outputs a posterior distribution and the amount of surprise generated by the new data. In our implementation, a chain of these surprise units is connected to each pixel in each level of each feature map such that the first unit in the chain receives its data directly from the detectors output value, as further described below. This data is modeled as a Poisson distribution $M(\lambda)$ (an assumption which is backed by recordings of the firing statistics of visual cortex cells), where the rate parameter λ is simply estimated as the detector's output value. Modeling the prior and posterior as Gamma distributions allows us to use the posterior from a previous step as the prior for the current step in a Bayesian learning configuration (Gamma is the conjugate prior on Poisson data). The Kullback- Leibler divergence is then used to measure the difference between these two distributions and this difference is labeled as 'Surprise.' In this way, the detector is able to constantly update its belief of the distribution of a single feature detector's response and to output a measure of the amount by which a new feature response violates that learned distribution.

By applying such a surprise detector to each feature map at each spatial scale, we are able to detect unexpected events of various sizes and types. However, to effectively eliminate repeating or periodic motions in the image (e.g., trees in the wind), it is also desirable to detect surprise at multiple temporal scales. In addition, while monitoring surprise over time may be useful to detect local transient events, computing surprise over space is also important to focus onto those events, which are also spatially salient. To accomplish this, we set up a network of these surprise detectors as shown in fig. 2 to form a complete temporal and spatial surprise computation unit. The input to this unit is a single feature map from a single pyramid scale. At each location in this map, we compute both the spatial and temporal surprise over five temporal scales. At the first temporal scale, the data from the feature map (shown with red arrows) is fed into the data inputs of a temporal surprise detector shown on the left, and a spatial surprise detector shown on the right. The temporal surprise detector uses the posterior from the previous time step as its prior and outputs the amount of surprise generated from this newest feature response. The spatial surprise detector uses the same feature response as input, but uses an average of the surrounding responses weighted by a Gaussian envelope as its prior distribution. This spatial surprise detector then outputs the novelty of a stimulus based only on the surrounding feature responses from the current frame. The posterior produced by this spatial surprise detector is ignored, but the posterior produced by the temporal detector is used to produce a pixel in the input map for the next temporal scale. Figure 3 shows the full system with surprise detectors implemented over the different feature channels, spatial scales, and temporal scales.

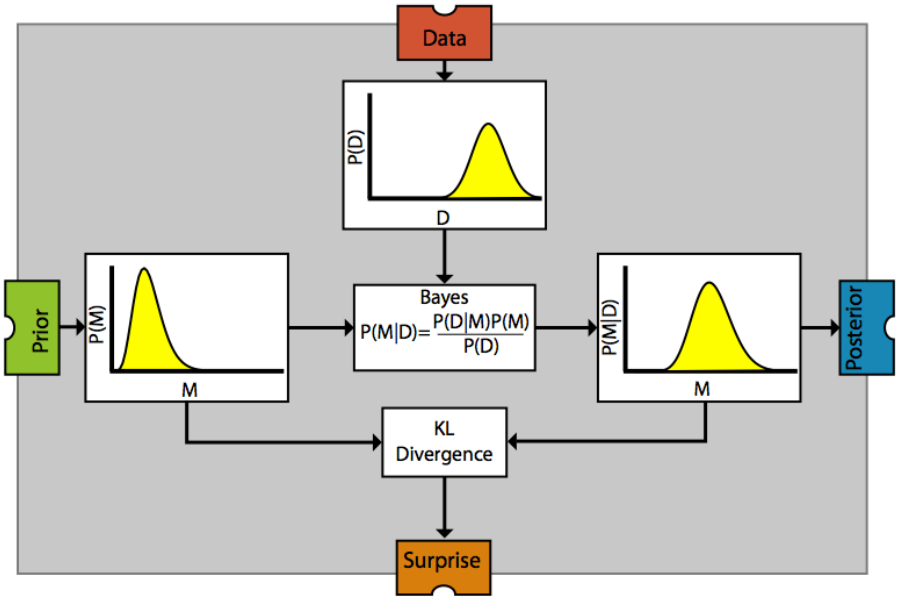


Figure 1: A single surprise unit which computes a posterior distribution from a prior and an observed data distribution and outputs a level of surprise proportional to the difference between the prior and posterior.

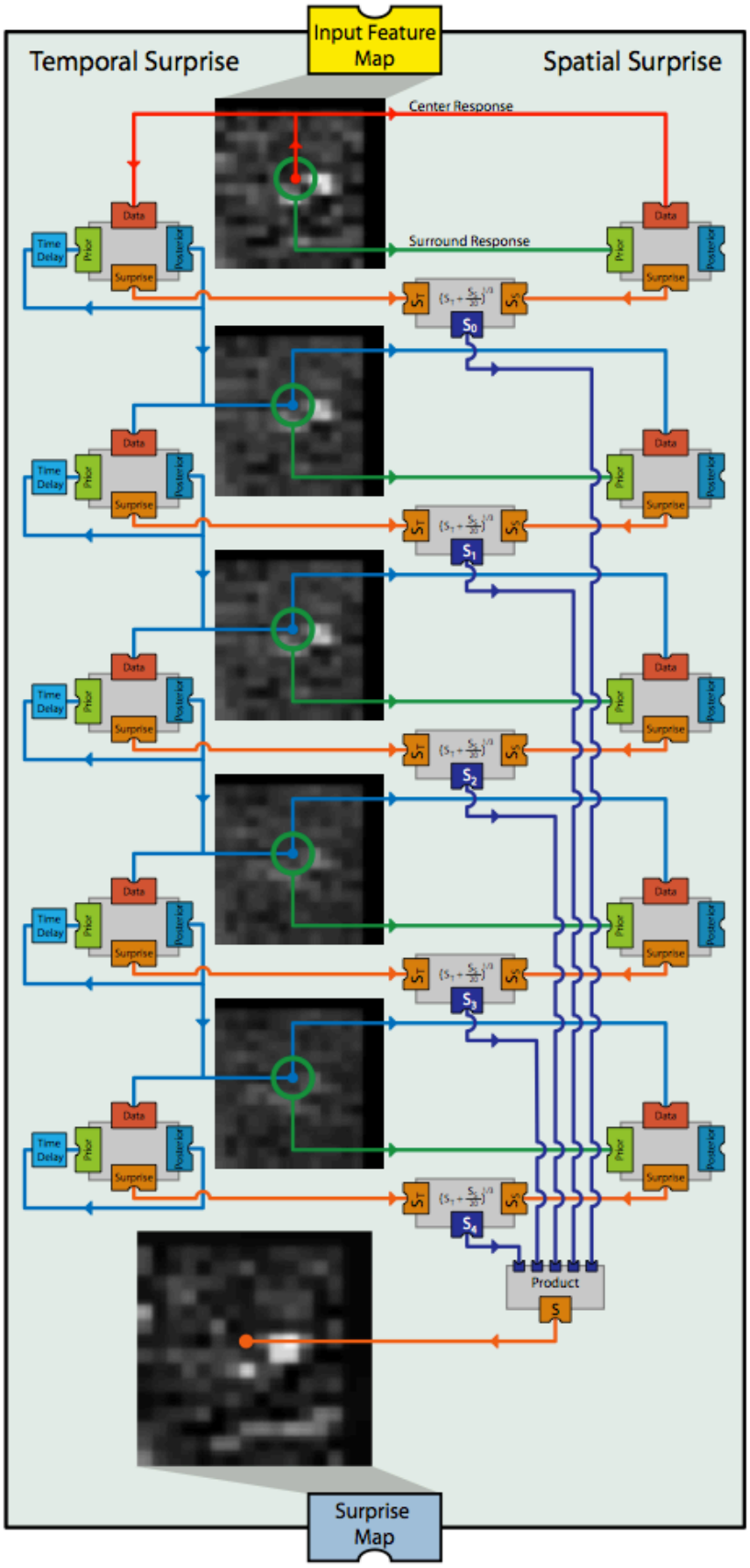


Figure 2: Surprise map computation unit which takes as input a new feature map from the current frame and outputs a map indicating the level of surprise at each pixel in the corresponding input map. Data path is shown only for a single pixel, but is replicated across all pixels in the implementation.

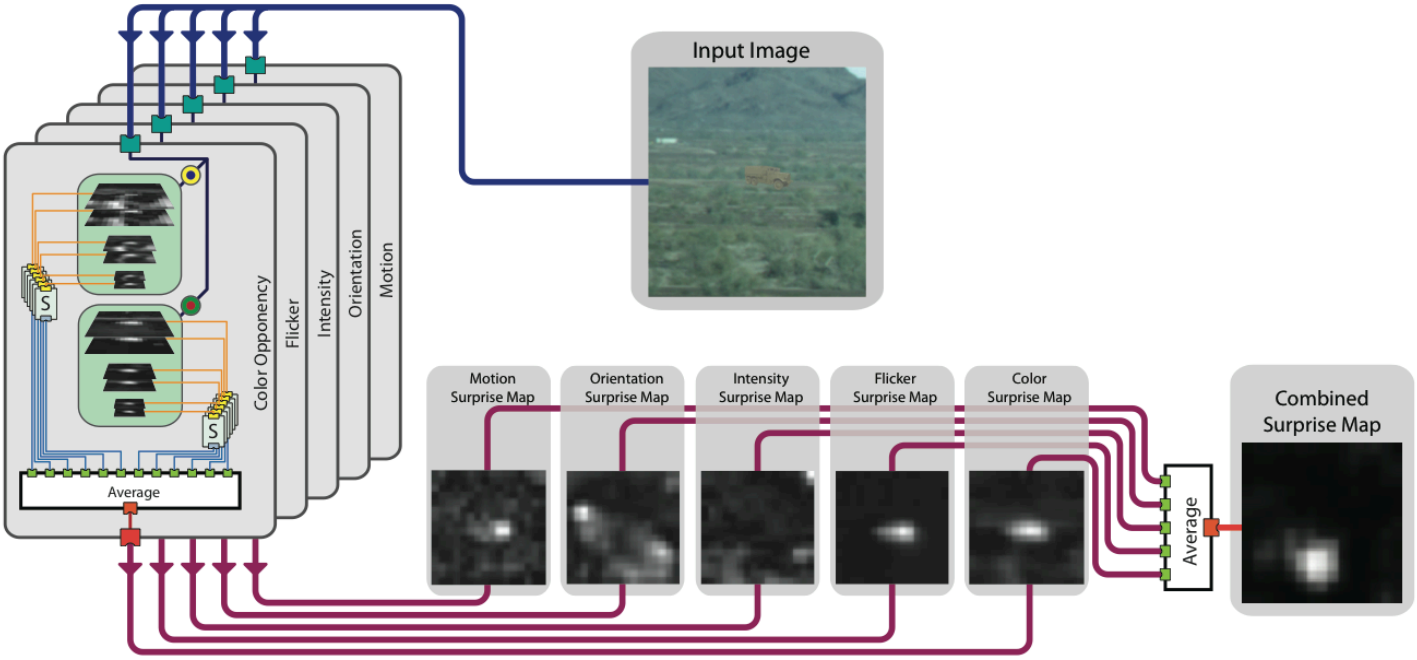


Figure 3: Full surprise algorithm. Surprise maps are computed for different feature channels (motion, orientation, intensity, flicker, and color) and combined to yield the final multimodal surprise map. Any location that reaches a surprise level above some threshold will yield a surprising event.

Testing and validation:

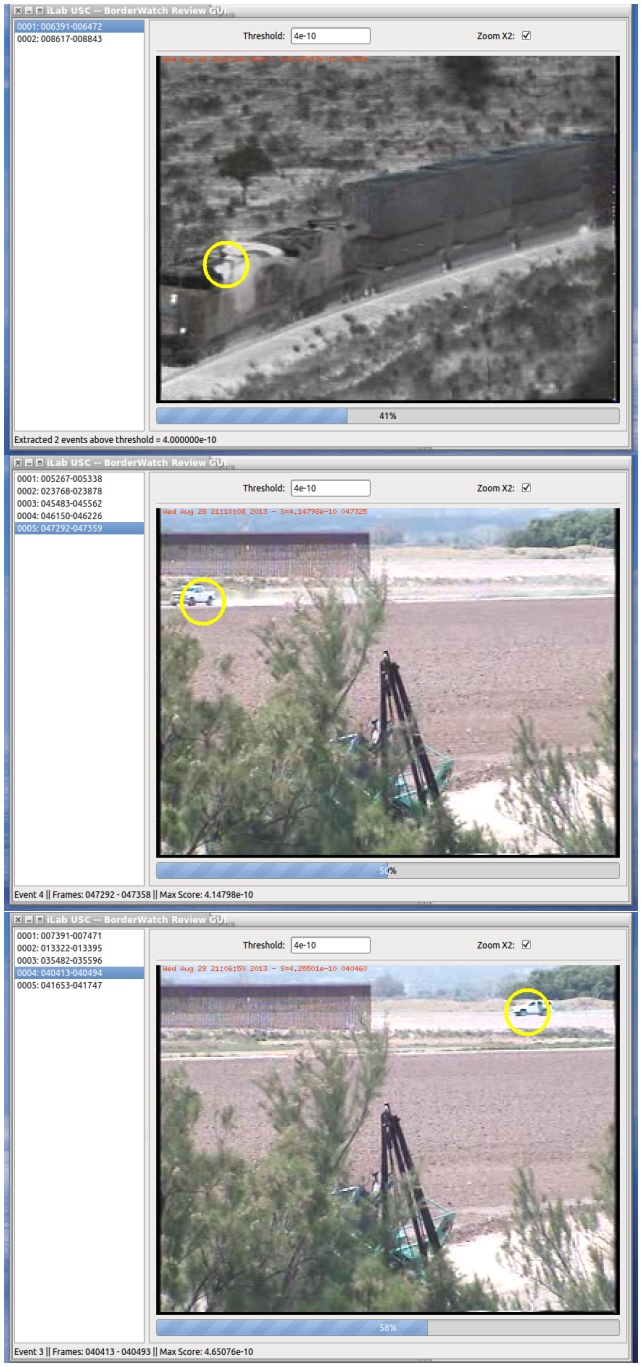
We applied our system to video clips from cameras along the Texas-Mexico border (blueservo project, which unfortunately went defunct shortly after we managed to grab 2,783 video clips from 24 different cameras, some RGB color and some infrared).

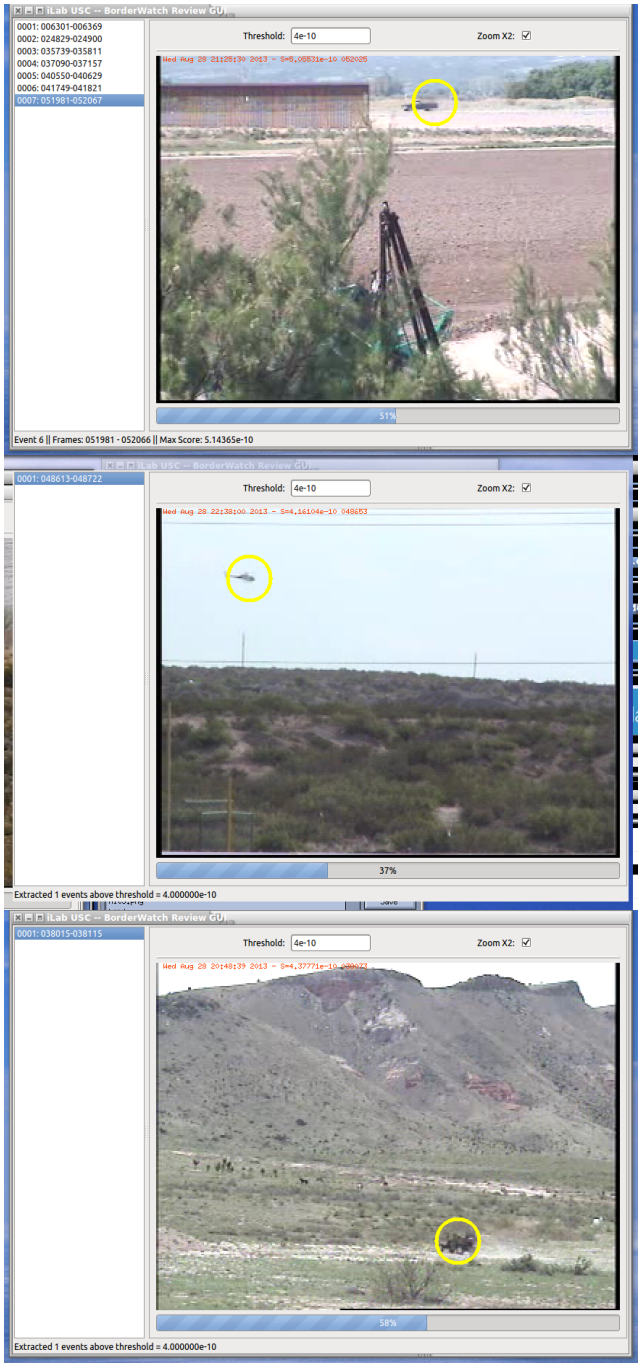
We first examined a smaller set of 89 videos (about 4.8 million frames and 89 hours of playtime). We set a fixed surprise threshold in our algorithm, such that any video frame where that threshold was crossed would be noted as a surprising event. In a graphical user interface, we displayed only surprising events, as short video clips of +/- 10 seconds around each surprising point. The surprise threshold was set once and for all before we ran the 89 clips, by manually looking for a reasonable value using other clips than the 89 tested.

Our algorithm correctly dismissed over 99% of all video frames as being boring (unsurprising), while correctly detecting 56 surprising events (later manually classified as 16 *interesting/relevant* events such as cars, boats, etc and 40 birds that flew through the field of view of the camera), with only 4 false alarms (camera glitches) and 2 misses (1 truck and 1 boat).

Examples of hits, birds, correct rejections, false alarms, and misses are shown in the following figures.

Overall, the algorithm appears highly performing for this application, with only 2 misses and 4 false alarms after processing 4.8 million video frames. The human workload was reduced by our algorithm from watching 89 hours of overall highly boring video to just 60 (events) x 20s (each event) = 20 minutes, i.e., a reduction of 270x (with 2 misses). In comparison, a system at chance level would only cut the viewing time by a factor 2 (45 hours, or 135x more viewing than with our algorithm), while incurring a 50% probability of miss (29 misses).





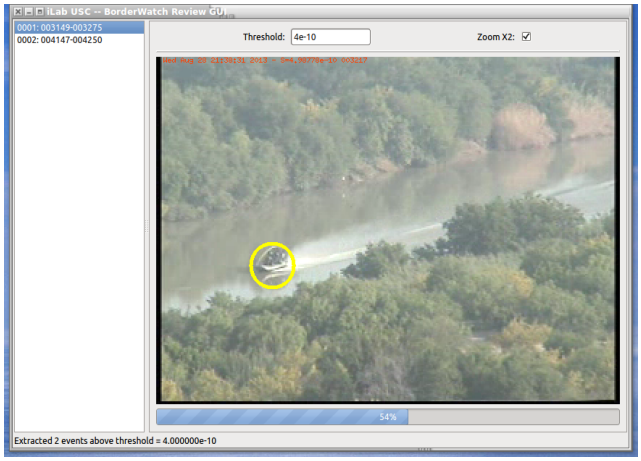


Figure 4: Example of surprising events detected by our algorithm (hits). From left to right, then top to bottom: A train; a truck; a truck; a helicopter; a truck; a dune buggy; a boat. Our review graphical user interface allows users to select a surprise threshold and then lists (left sidebar) the events that are above threshold. The user can then click on each event in turn, at which point a short video of +/- 10 seconds around the event is played. In our experiments, the threshold was fixed to a single value (4.0 Wows of surprise) chosen by examining other clips than those processed here, for all cameras and all clips.

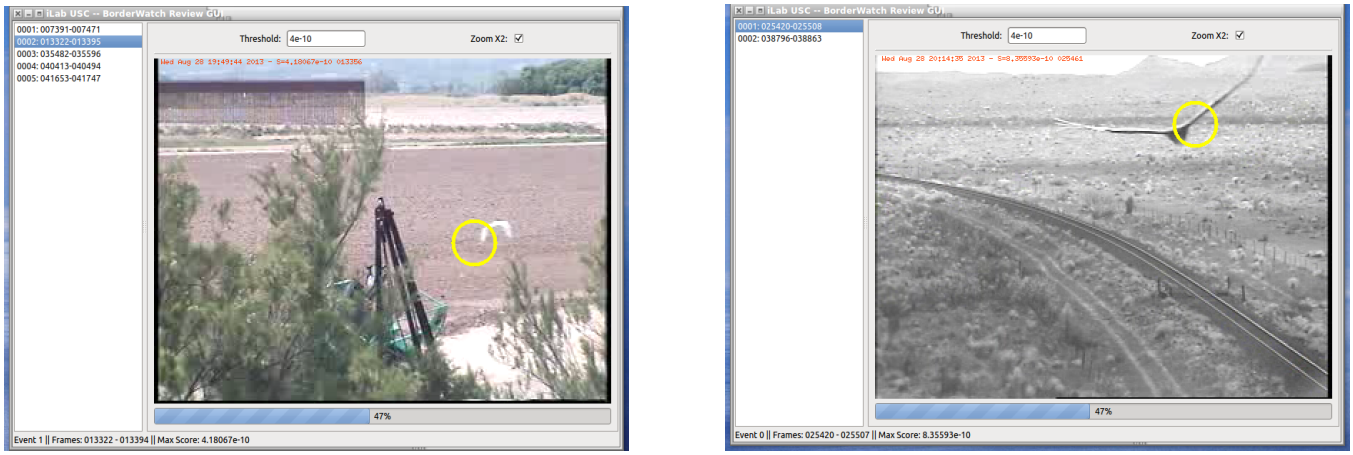


Figure 5: Example of birds. As far as being surprising, those may be considered hits, as they definitely catch the eye when they occur. In years 2 and 3 we will study in more depth how to eliminate surprising events that are uninteresting or irrelevant with respect to the task at hand (here, surveillance of the border with respect to illegal human crossing).

Figure 6: Examples of correct rejections from several cameras. Nothing was detected as significantly surprising by our algorithm in these videos. (Note how we here selected a lower surprise threshold in our software to display those images, as otherwise they had been suppressed as containing nothing surprising).

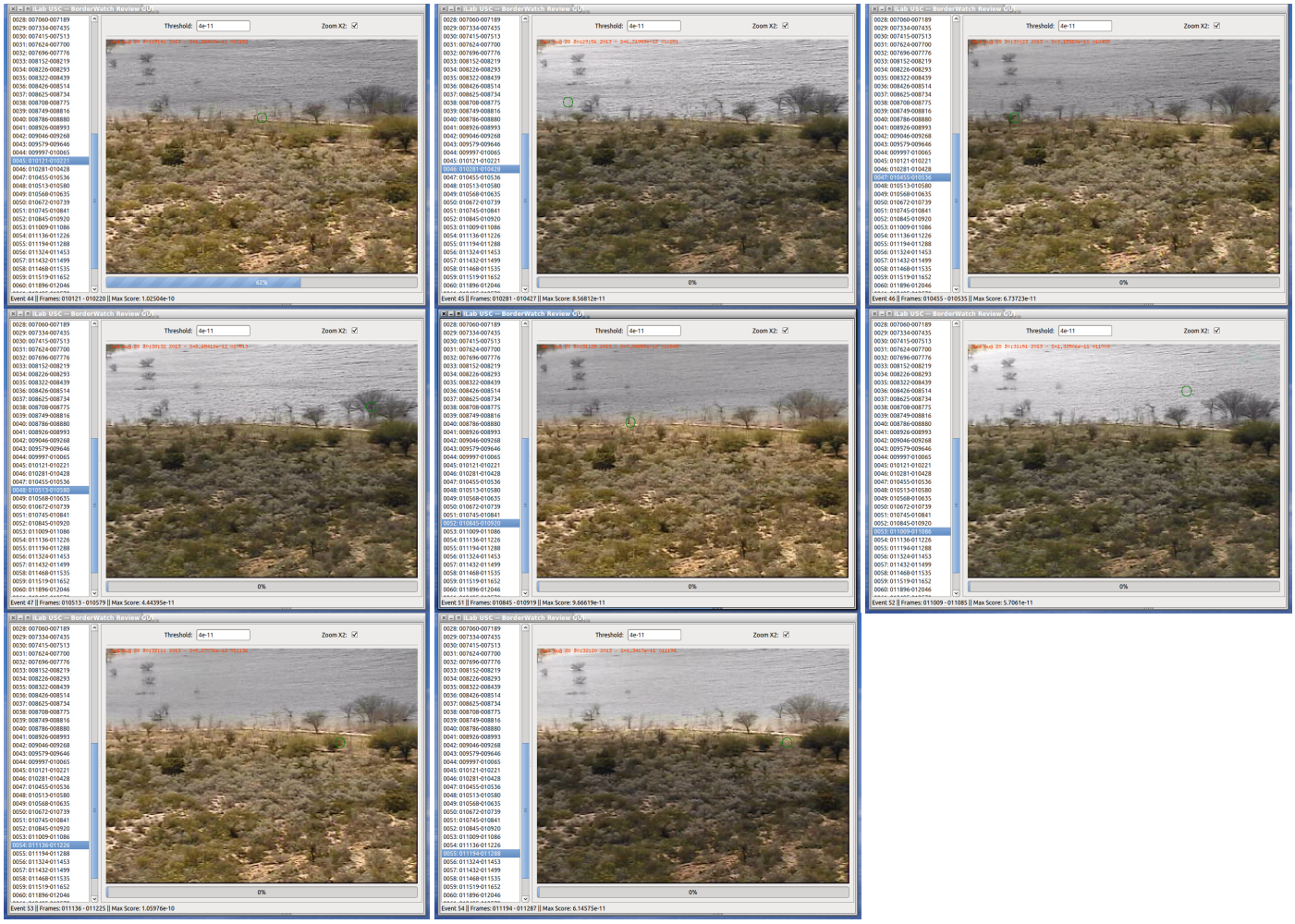


Figure 7: Example of correct rejection of an entire clip. Nothing was detected as surprising in this hour-long video clip, although we can observe significant variations in colors, illumination, water ripples, etc throughout the duration of the video clip. These variations were correctly dismissed by our algorithm.

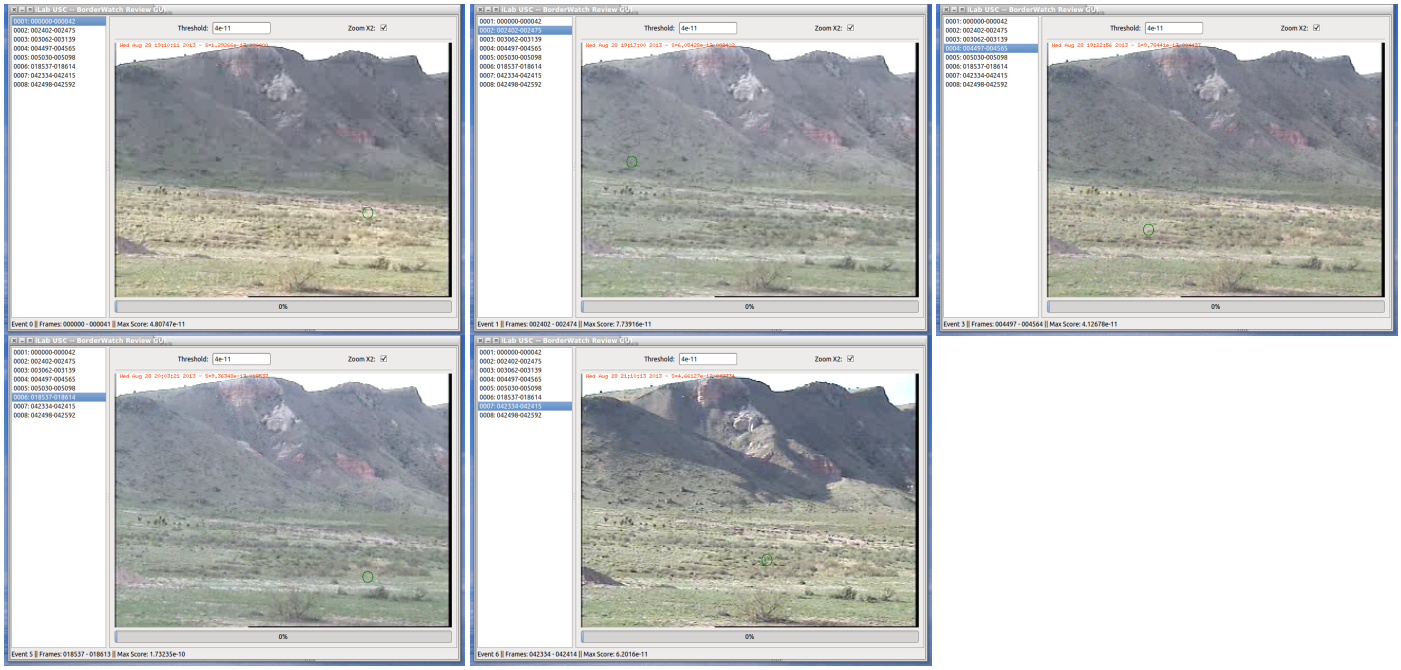


Figure 8: Another example of correct rejections. Nothing surprising was detected by our algorithm despite significant changes in color, illumination, and shadows due to cloud and sun movement.

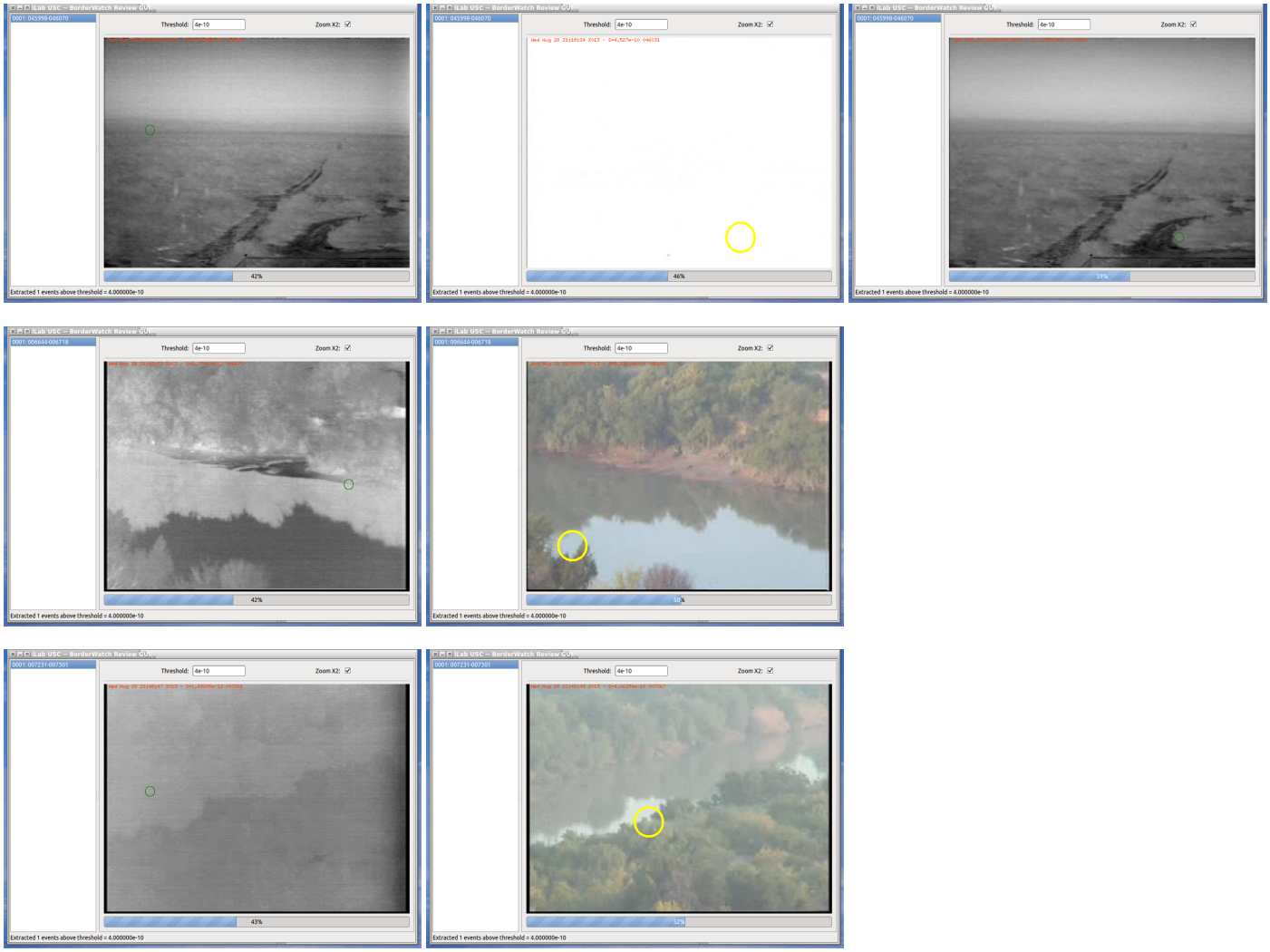


Figure 9: Examples of false alarms. Top row: In one clip, the image turned solid white for a brief moment, generating a high surprise event in our algorithm. Middle and bottom rows: In 3 other clips (two shown, one each row), the camera switched from night-vision mode to daylight color mode, creating high surprise at that transition point.

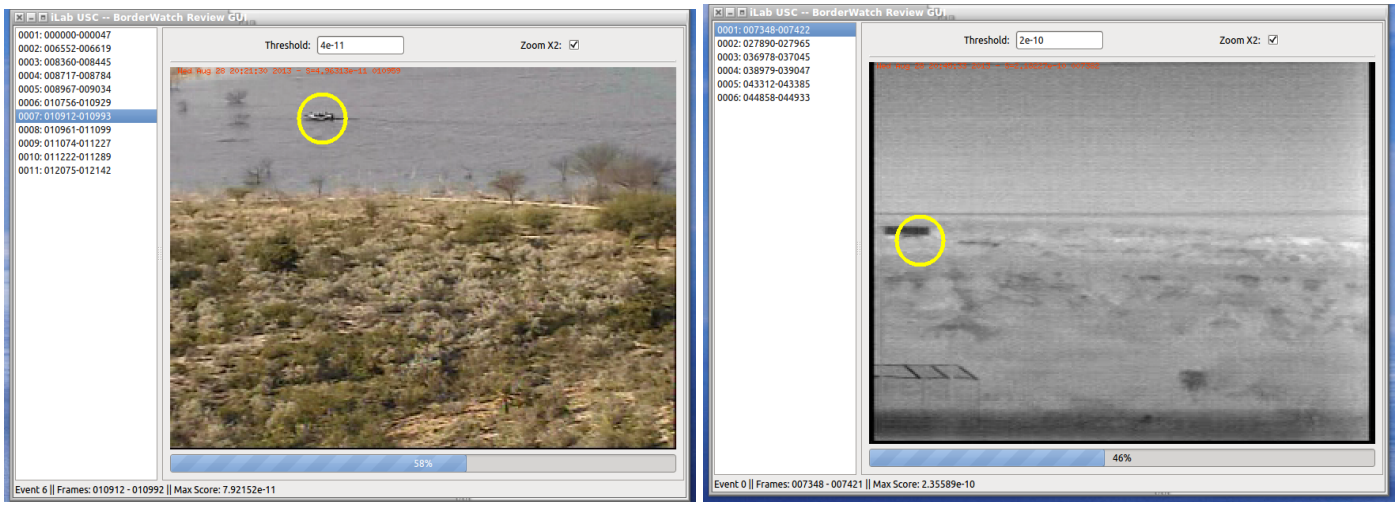


Figure 10: Example of misses, which we display here by setting a lower surprise threshold in the software. Left: a boat; right: a truck. In the left camera, framerate was lower than in the other cameras, and thus the boat appeared to move very slowly and was dismissed as boring. Simply setting a lower surprise threshold for that specific camera fixed the problem, without generating additional false alarms. Likewise, the right camera is an infrared one, which might benefit from using a lower surprise threshold as well, since it has no color information.

2. Modeling top-down attention to relevant items in the world

As seen in our surprise system above, and as argued in the proposal for this project, sometimes surprising events are not relevant to the task or interesting: in the examples above, we detected 40 surprising birds flying into the cameras' field of view. While those indeed are surprising and could startle a human operator, they are not relevant to the problem of border control.

We investigating top-down attention control towards task-relevant items, and wrote two review-style textbook chapters in the process (Itti & Borji, 2013; 2014).

This has further allowed us to clarify several concepts in the attention literature, with respect to (1) the role of objects in guiding attention, as opposed to attention being driven by simpler low-level visual features (Borji et al., 2013 JOV), and (2) the difference between low-level saliency and interest, as better distinguished by our new measure of explicit saliency (Borji et al., 2013 VR).

Below we report a model of top-down attention that directly attacks the problem of finding the most relevant information in a video stream, given a particular task definition.

We created a new framework to model top-down overt visual attention based on reasoning, in a task-dependent manner, about objects present in the scene and about previous eye movements. We designed a Dynamic Bayesian Network (DBN) that infers probability distributions over attended objects and spatial locations directly from observed data. Two basic concepts in this model are 1) taking advantage of the sequence structure of tasks, which allows to predict the future fixations from past fixations and knowledge about objects present in the scene. Graphical models have indeed been very successful in the past to model sequences with applications in different domains, including biology, time series modeling, and video processing, and 2) computing attention at the object level. Since objects are essential building blocks in scenes, it is reasonable to assume that humans have instantaneous access to task-driven object-level variables (as opposed to only gist-like, scene- global, representations). Briefly, the model works by defining a Bayesian network over object variables that matter for the task. For example, in a video game where one runs a hot-dog stand and has to serve multiple hungry customers while managing the grill, those include raw sausages, cooked sausages, buns, ketchup, etc. (Figure 11). Then, existing objects in the scene, as well as the previous attended object, provide evidence toward the next attended object (Figure 11). The model also allows to read out which spatial location will be attended, thus allowing one to verify its accuracy against the next actual fixation of the human player. The parameters of the network are learned directly from training data in the same form as the test data (human players playing the game). This object-based model was significantly more predictive of eye fixations compared to simpler classifier-based models, also developed by the same authors, that map a signature of a scene to eye positions, several state-of-the-art bottom-up saliency models, as well as brute-force algorithms such as mean eye position (Figure 12). This points toward the efficacy of this class of models for modeling spatio-temporal visual data in presence of a task and hence a promising direction for future. Probabilistic inference in this model is performed over object-related functions which are fed from manual annotations of objects in video scenes or by state-of-the-art object detection models. The same model has been successfully applied to different types of tasks (driving, flight combat, etc; Figure 13), demonstrating generality of the approach.

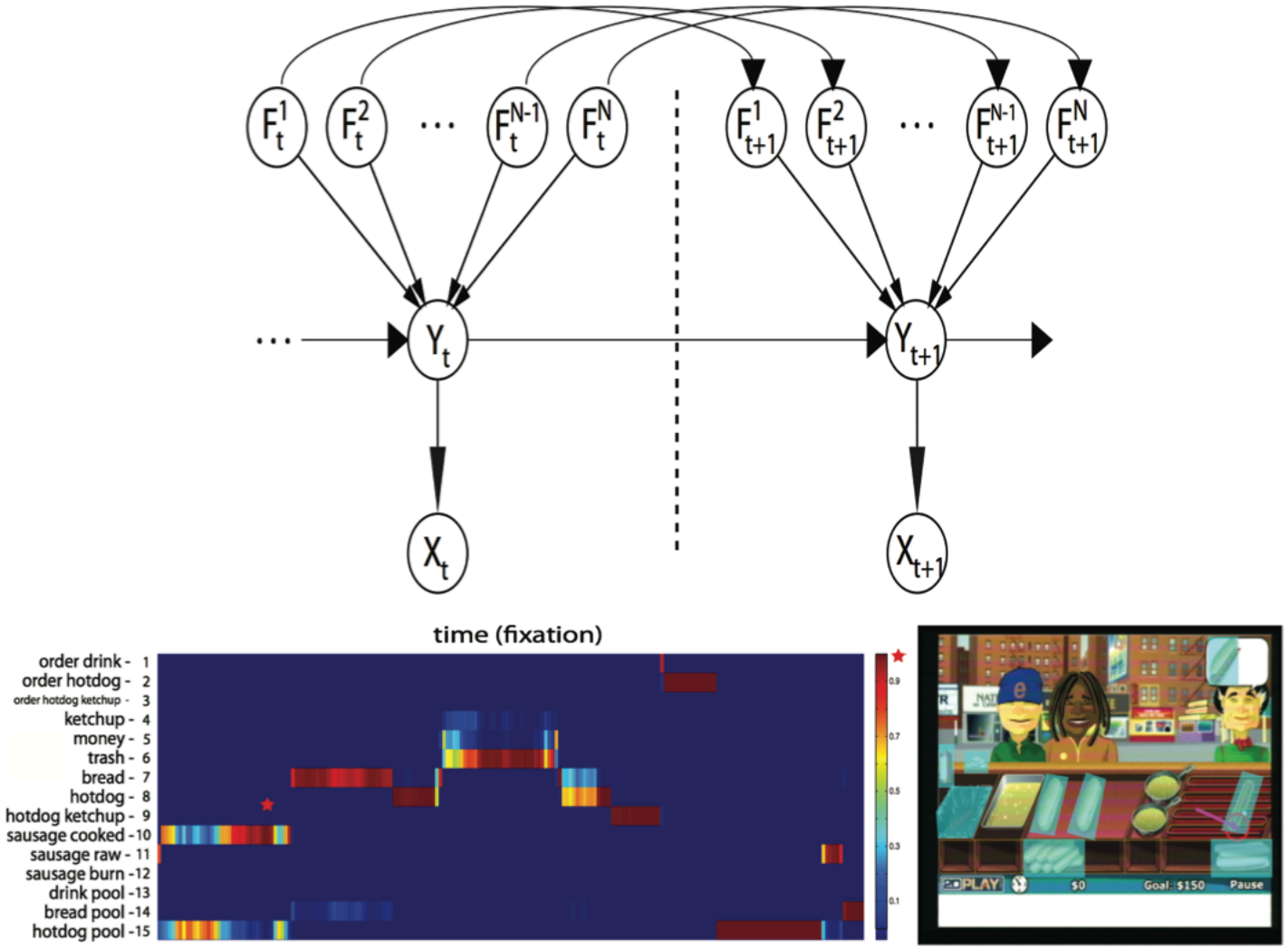


Figure 11: Graphical representation of our DBN approach unrolled over two time slices. X_t is the current saccade position, Y_t is the currently attended object, and F^t_i is the function that describes object i in the current scene. All variables are discrete. Also shown is a time series plot of probability of objects being attended and a sample frame with tagged objects and eye fixation overlaid.

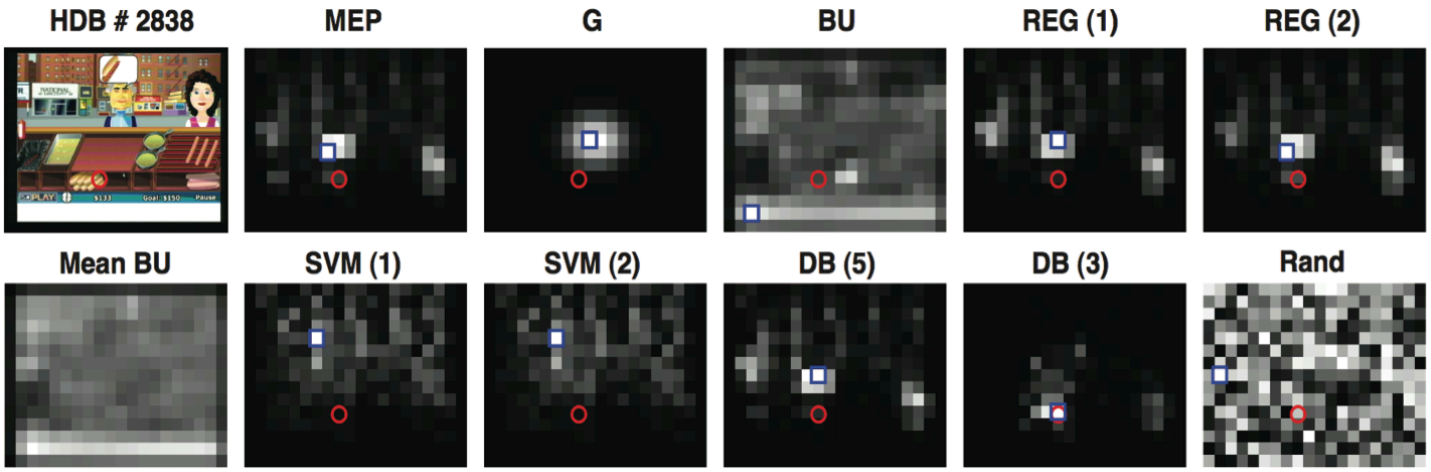


Figure 12: Sample predicted saccade maps of the DBN model. Each red circle indicates the human observer's eye position superimposed with each maps peak location (blue squares). Smaller distance indicates better prediction. Images from top-left to bottom-right are: a sample frame from the hot-dog bush game where the player has to serve customers food and drink, MEP stands for the mean eye position over all frames during the game play, G is just a trivial Gaussian map at the image center, BU is the bottom-up saliency map of the Itti model, REG(1) is a regression model which maps the previous attended object to the current attended object and fixation location, REG(2) is similar to REG(1) but the input vector consists of the available objects at the scene augmented with the previously attended object, SVM(1) and SVM(1) correspond to REG(1) and REG(2) but using an SVM classifier, Mean BU is the average BU map showing which regions are salient throughout the game course, Similarly DBN(1) and DBN(2) correspond to REG(1) and REG(2) meaning that in DBN(1) network slice consists of just one node for previously attended object while in DBN(2) each network slice consists of the previously attended object as well information of the previous objects in the scene, and finally Rand is a white noise random map.

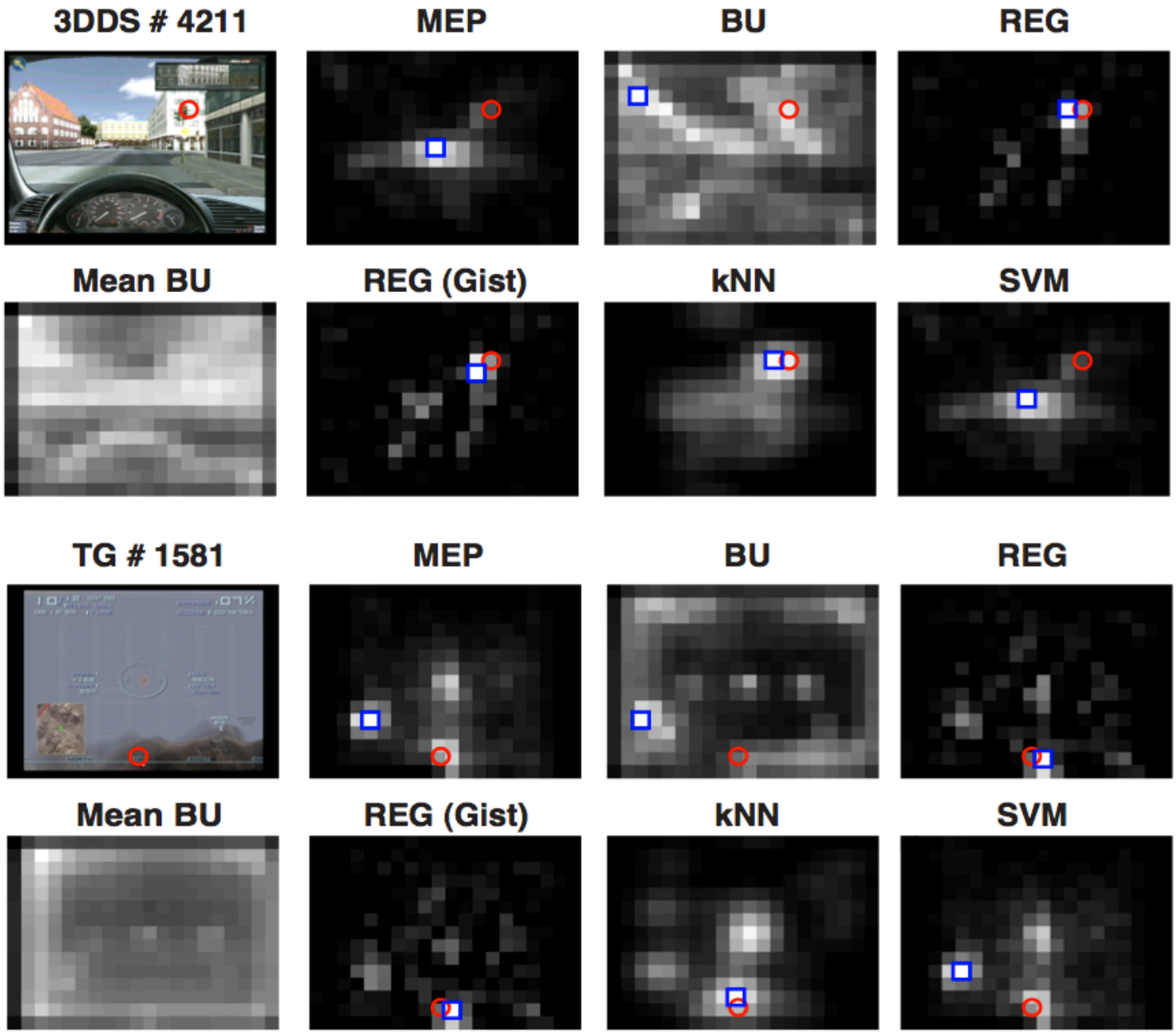


Figure 13: Applications of our top-down model to driving (top two rows) and flight combat (bottom two rows). Abbreviations are as in Fig. 12.

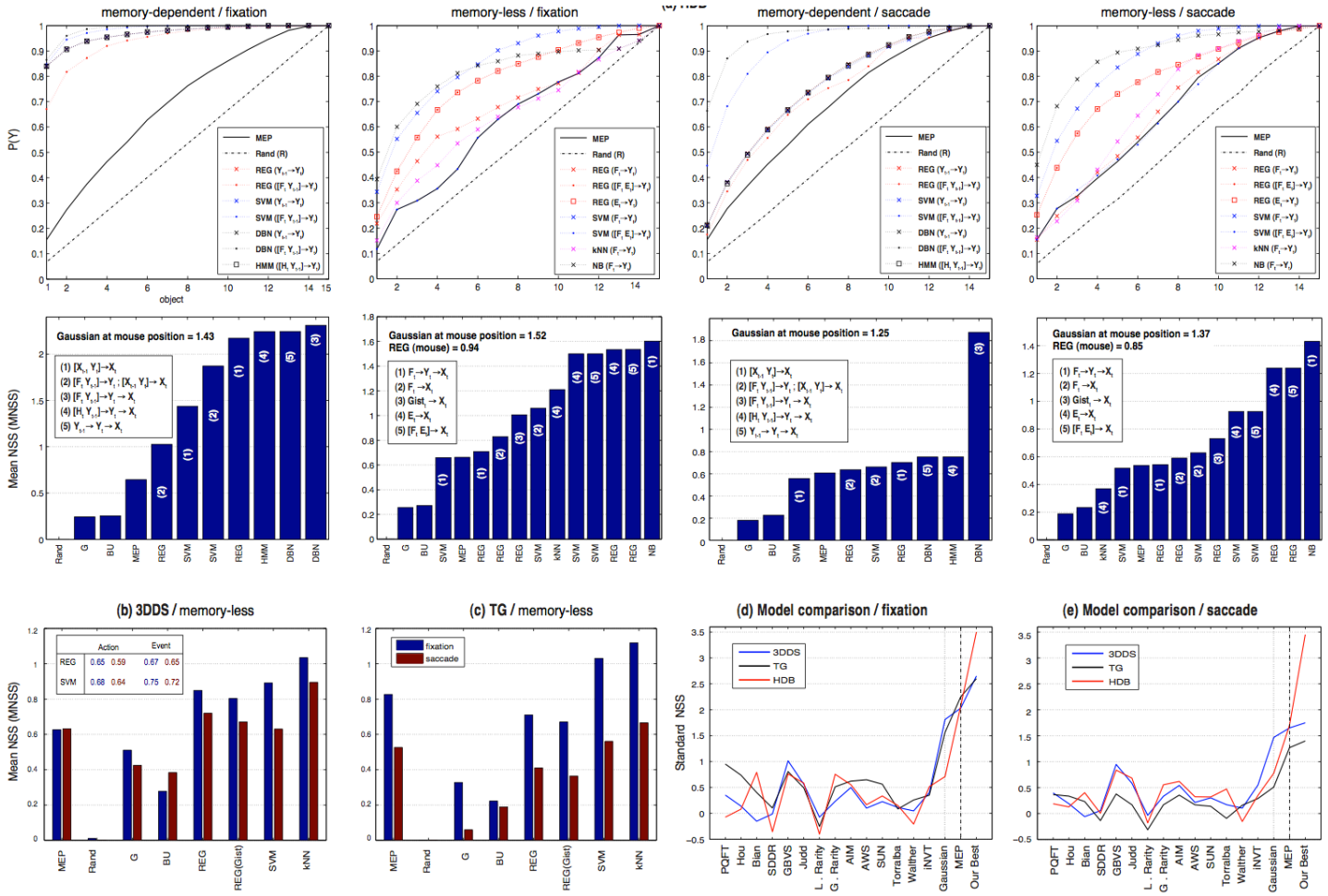


Figure 14. Gaze prediction accuracies. a) probability of correctly attended object (first row) and MNSS scores (which measure the extent to which an observer looked at model-predicted locations above chance level) for prediction of saccades and fixation positions (second row) for all models. White legends on bars show the mapping from feature types to gaze position X. For instance, REG (Ft → Yt → Xt) maps object features to the attended object and then maps this prediction to the attended location using regression. Property functions f(.) in HDB (hot-dog stand) indicate whether an object exists in the scene or not (binary). b) and c) MNSS scores of our classifiers over 3DDS (driving) and TG (flight combat) games, d) and e) NSS scores (corresponding to $\gamma = 0$ in MNSS) of bottom-up models for saccade prediction over 3 games. Almost all of bottom-up models perform lower than MEP and Gaussian, while our models perform higher. Some models are worse than random (NSS < 0) since saccades are top-down driven instead of bottom-up.

This model for the first time provides a fully computational approach to determining what is the next most top-down relevant location one should look at while engaged in a complex task, and it predicts human behavior better than other attention models.

3. Defining and implementing goal relevance

In the second year, we focused on two major thrusts: a new mathematical definition of goal relevance that naturally combines with our work on surprise, and applications to tasks such as analysis of eye movements, path planning, and navigation. Our definition of goal relevance stems from first principles. Assuming an agent with one or more goals, we consider the agent's distribution of beliefs over all possible ways it could achieve its goal(s). When new information is received from the agent's sensors, this distribution is updated (e.g., sensing a new obstacle may decrease the agent's belief that certain paths may lead to the goal). We then measure the rele-

vance of the observation as the Kullback-Leibler divergence between the prior and posterior belief distributions. Thus, data observations that does not affect the agent's beliefs over how it will reach its goal are deemed irrelevant, while data observations that force the agent to significantly re-estimate how it might achieve its goal is relevant. To test our proposed definition, we conducted human experiments to gauge the human intuitive notion of what makes some information more relevant. We found that our proposed approach matches the intuitive human notion of relevance, significantly better than alternative approaches. We applied these concepts to the analysis of eye movements during complex tasks and to robotic path planning and navigation. Our new proposed definition of relevance is general and derived from first principles, and for the first time provides a quantitative measure of relevance that is directly applicable to many domains (as opposed to previous definitions, often restricted to specific domains, such as web search, linguistic analysis, etc).

Although many studies have investigated the effects of tasks and goals on human attention, no model to date has provided a general quantitative definition and metric of these effects. Most studies of goal-oriented attention either have been qualitative and descriptive, or have used computational models that implicitly learn from examples how participants deploy attention under various tasks. Here we propose a theoretical definition for the information value of data observations with respect to a goal, which we call goal relevance. We consider the probability distribution of an agent's subjective beliefs over how a goal (or set of goals) could be achieved (e.g., beliefs over the set of all possible paths from start to finish in an obstacle course). When new data is observed, the belief distribution is updated (e.g., a newly detected obstacle may invalidate some paths). We then simply measure the goal relevance of that observed data as the Kullback-Leibler divergence between belief distributions before and after the observation. To test our definition, we compare its predictions to responses of 35 human participants, who were presented 200 times with two side-by-side images of a simple 2D obstacle course, with a different new obstacle added to each image. They indicated which of the two new obstacles was more relevant to the task of traveling from start to finish. Our definition agreed with participant responses (85% correct), slightly better than inter-observer agreement (82%), and significantly better than 3 competing models based on machine learning classifiers (74% for best tested). Our new definition of goal relevance is general, quantitative, explicit, and allows one to put a number onto the previously elusive notion of relevance of observations to a goal.

Relevance is a concept that humans use all the time. It is a measure of the importance of observations with respect to a context. This allows humans to prioritize and filter sensory input by directing their attention. While driving, for example, people will often ignore things going on inside the car, in the sky, or in the oncoming lane. This might be accomplished with heuristics learned from experience, which humans do well [6].

Relevance has been studied extensively in linguistics, document search, and data mining [5]. In these cases, it is usually formulated as a statistical similarity metric between two concepts. This works for explaining why ”Obama” is relevant to ”politics”, or ”E. coli” is relevant to ”disease”. However, these similarity metrics do not naturally extend to the driving example, because driving is a task. During driving, the relevance of objects is not determined by their similarity to another object, but by how they affect the driver's goal of safely reaching a destination. This type of relevance, which will be referred to as goal relevance, has not received much attention in computer vision, robotics, or perception, even though it is useful for perception in humans [4].

Many studies (see below) have shown that task influences human attention [8], providing a wealth of qualitative and empirical data. Attempts to model this quantitatively typically establish correlations between computed features and the tasks being investigated [17, 16]. Our motivation is towards the development of top-down attention models that are based on theoretical knowledge. Attention during a task can be seen as a search for relevant data. Thus, in the hopes of better understanding the effects of task on attention, the present study investigates the manner in which humans decide what input is considered relevant.

3.1. Background

In the information retrieval community, there is still debate over the precise definition of relevance. Hjorland discusses five different views of relevance and argues for what he calls the ”subject knowledge view”, which

suggests that relevance is not an inherent attribute but rather is dependent on the knowledge or beliefs of the subject evaluating the relevance [9]. All five of the presented views focus on relevance as applied to web search, and Hjorland does not focus on the case where an agent is pursuing a goal. He does refer to one of his previous papers, in which he provides an interesting definition of relevance:

> *Something (A) is relevant to a task (T) if it increases the likelihood of accomplishing the goal (G) which is implied by T [10].*

This is fairly intuitive, but is still vague for use in obtaining a quantitative result. It requires a likelihood function specific to the goal G which must capture most of the problem. There exists a wide range of types of goals, from navigation to purchasing groceries to solving math problems, so the computation of relevance will certainly be specific to each goal. However, a definition that more clearly defines the computation would be preferable.

It has been well established that tasks influence human attention, ever since the studies of Buswell [3]. Hayhoe and Ballard provide a review of recent advances, in which they note that attention has been examined (among other things) in a number of visuo-motor tasks including walking, driving, sports, and making sandwiches [8]. In these cases, the task is usually considered a top-down influence on attention. In one study, it was demonstrated that local and global visual features around subjects' focal point of attention were sufficient to predict the task being performed [17]. Another [14] showed that top-level features about the task affected the saccade pattern. Navalpakkam and Itti [15] develop a model for predicting saccades, using a "long term memory" to represent top-down task information, containing abstract knowledge initialized by hand and updated via learning. Several studies have used learning from examples to implicitly capture task influences on eye movements [16, 2]. Despite the success of these studies, we still do not understand algorithmically how a given task specifically affects attention. Thus, we do not have models that, given a task, can predict *a priori* which objects will attract more attention. This provides motivation for the present study.

Gottlieb and Balan [7] present an experiment on attention and decision-making in monkeys that cannot be explained by any previous models. The authors suggest that attention is a decision process based on the utility of information. They also point out neurobiological evidence that the same parietal neurons carrying information about saccadic motor decisions also carry information unrelated to any other aspects of those decisions. It is proposed that this additional information reflects the information utility, which then affects the saccades. The question of how such a utility metric is computed is left open. The present study attempts to address this.

3.2. Defining Goal Relevance

3.2.1 Theory

What makes a data observation relevant to a goal? Let us examine our intuition using the driving example. Seeing a car brake in front of you or knowing about the movement of objects on the road is certainly relevant to the goal of safely reaching a destination. Hearing that a bridge is out seems extremely relevant, if that bridge was on one of the possible routes. However, learning that "Google stock is down 0.1%" isn't relevant at all. How can we support these intuitions while being abstract enough to apply to other types of goals?

To answer this question, we first identify what is common to all goals. First, since we are considering the relevance of data observations while pursuing a goal, there must be an agent that does the pursuing. Also, the act of pursuing a goal implies movement through some real or imaginary state space, from a current (start) state to one or several desired (goal) state or states. Note that like in standard search in Artificial Intelligence, the goal may be implicitly defined (any state that satisfies a number of requirements that establish goal test would qualify as goal state). We can define S as the set of all paths through this space which begin at the start and terminate at the goal. The job of the agent is to follow one of these paths (or a collection of paths that overlap near the current state of the agent), while monitoring the environment for possible new data observations that change S and that thus may warrant that the followed path or paths should be updated. Looking back at the intuitions above,

we can note that more relevant pieces of data may be the ones that cause larger changes in *S*. Indeed, if a data observation does not change *S*, it does not affect the agent's choices about how to reach the goal(s) and therefore it should not be considered relevant to the goal(s). If a data observation instead renders some or many of the current possible paths infeasible, or opens up one or more new possible paths, then it should be considered very relevant according to our proposed definition. This inspires our key idea, which is to define goal relevance by the amount of change in *S* caused by the new data observation.

We thus define the goal relevance of a data observation *D* with respect to an agent's probability distribution *S* over possible ways it could achieve its goals as a distance measure, $d(\cdot, \cdot)$, between the prior distribution of beliefs $P(S)$ and the posterior distribution $P(S|D)$ after observation of data *D*:

$$Relevance(D, S) = d(P(S|D), P(S)) \tag{1}$$

While many measures are available for $d(\cdot, \cdot)$, here, inspired by the related manner in which the concept of surprise was mathematically defined by Itti & Baldi [12] in a rigorous, quantitative manner, we will use for numerical applications the Kullback-Leibler divergence $KL(\cdot, \cdot)$, such that:

$$Relevance(D, S) = KL(P(S|D), P(S)) = \int_S P(S|D) \log \frac{P(S|D)}{P(S)} dS \tag{2}$$

(Note however that the Kullback-Leibler divergence is not strictly a distance measure, as it is not symmetric; one could use the symmetric version $KL_{sym}(P,Q) = \frac{1}{2}(KL(P,Q) + KL(Q,P))$ interchangeably). Our definition thus interprets goal relevance as the degree to which the data observation *D* yielded surprising change (in Itti and Baldi's terms) in the observer's beliefs over how it thought it could achieve its goal(s). Therefore, in a deterministic universe where the agent also has perfect information, nothing is relevant because the agent can optimally plan how to accomplish its goal just based on the initial *S* and while ignoring all sensory input. Relevance is important when there is uncertainty or there are environmental changes that cannot be predicted by the agent.

3.2.2. Implementation

Ideally, the proposed definition of goal relevance would be compared against the prevailing one. However, to the authors' knowledge, there are no other quantitative models that explain the relevance of observations to goals. Instead, since our definition attempts to explain human cognition, the above model is used to predict human responses to a relevance task. Given a 2-D environment with obstacles, a starting location, and a goal location, we asked participants to indicate which of two new obstacles added to the environment was the most relevant to the task of traveling from start to goal.

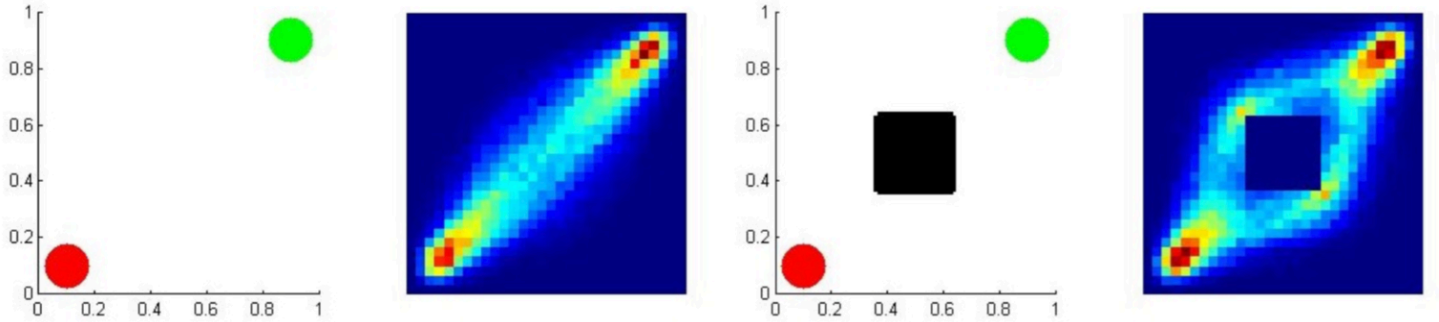


Figure 15: First and third: A simple 2-D environment without (first) and with (third) an obstacle being evaluated. Second and fourth: Probability distributions over the relevance of each grid cell, computed from RRT-based path sampling on the environments and Eqn. (2). Hotter locations are determined as more relevant to the task of traveling from start to goal, according to our definition of goal relevance.

To apply the equation for relevance, a model space must be chosen that can represent the possible paths. The 2-D environment space was used for this purpose. The space was discretized into a grid, and a path was represented by counting the number of path points that fell into each grid space. To construct prior and posterior path distributions related to adding one obstacle to the environment, 1000 paths were randomly sampled on the environment without the obstacle in question (prior), and another 1000 on the environment with the extra obstacle (posterior). Sampling was done by using Rapidly-exploring Random Trees (RRTs) [13] to repeatedly find a random path from the start to the goal. The grid counts from the resulting paths were added together and then normalized to create the distributions. This process is visualized in Figure 15, which shows a very simple environment and the resulting distributions. Once the prior and posterior distributions are known, the relevance equation can be applied directly to compute the relevance score associated with adding one obstacle to the environment.

To compare relevance between objects, the above is performed for each object, and the object with the largest relevance score is the most relevant to the goal. Note that when comparing multiple obstacles, the environment used to compute the prior distribution contains none of them, and so is the same for all obstacles. The posterior is computed using only the obstacle in question.

This implementation was used to compute relevance scores for both objects in the same image pairs as used for the human experiment, which we used as a model prediction of the human responses. The details of these image pairs are covered in the next section.

3.3. Experimental Methods

To investigate the human intuition behind goal relevance, we recruited human participants and asked them about the relative task relevance of objects. Participants were presented with image pairs, such as the ones shown in Figure 16. The images represent two dimensional environments, with randomly placed objects and starting/ending locations (yellow/green dots, respectively). In all cases, each image in the pair was identical to the other, except for one additional object in each image which was colored red. Subjects were told to imagine that they were walking across the “room” from the start location to the end location. They were then instructed to consider, “Which of the two red objects is more relevant to your goal? In other words, which would you pay more attention to, or be more concerned with, as you cross the room?” The subjects responded to each image pair by indicating whether they thought the additional object in the image on the left or the right was more relevant. Response time was recorded along with this decision.

Figure 16: Left pair: Image pair with rectangular objects. Right pair: Image pair with convex polygon objects. The first and third images contain the more relevant new obstacles according to our theory. We thus expected that a majority of human responses would be "Left" for both image pairs.

Each subject was presented with 200 image pairs. The first 100 contained only rectangular objects, and the second 100 contained objects that were convex polygons. This allowed us to investigate if shape had any effect on goal relevance. All participants viewed the same 200 pairs, which were randomly generated in advance of the study. For each participant, presentation on the left or right for the two images in each pair was randomly shuffled, as was the presentation order of the image pairs. However, the rectangular-object image pairs always came prior to the convex-polygonal pairs. To increase the number of interesting image pairs (i.e., pairs in which both objects could be considered relevant), image pairs were only used if both of the additional objects intersected or lay within the circle whose diameter was the line between the start and end points.

In total, 38 undergraduate students participated in our study. Subjects had normal or corrected-to-normal vision, and were compensated with course credits. The experimental methods were approved by our university's Institutional Review Board (IRB).

The human responses to the image pairs were compared against the predictions of several models. First, they were compared to the goal relevance model described in this paper. In addition, three Support Vector Machines (SVMs) were trained using different features to provide benchmarks for the model. SVM1 was trained using two features, representing the areas of the two objects in question; it should predict human responses very well if participants preferentially picked the largest obstacle as being more relevant. SVM2 used an additional two features, the distance from the midpoint between start and goal to the centroids of the two objects; this model thus also coarsely accounted for how close the added obstacles were to the straight line from start to goal. SVM3 used the same two distance features and also a third feature, the difference in optimal path length between the two images; thus, this model should be able to capture whether humans strongly based their decision on how adding an obstacle might affect the optimal route from start to goal. The area features were not included in SVM3 because they decreased the performance. Several other features were experimented with, such as distances from start/goal points and obstacle circumferences. The features used in SVM3 were chosen because they provided better performance than in any of our other attempts. This is discussed further in the Results section.

In all cases, the SVMs were trained and tested using a leave-one-out scheme. Each image pair was tested using an SVM trained on the other 199 pairs. This allowed us to maximize the use of our data. It also gave the maximum available amount of training data to these models, while our proposed definition of relevance used no training at all (just Eqn. (2)).

3.4. Results

Data from 3 subjects was excluded due to drastically different responses over the 200 pairs compared to the majority of all other subjects. These subjects may have been selecting the least relevant object, rather than the most. The data from the remaining 35 subjects was analyzed to produce our results.

For each image pair, the number of responses for the left and the right object were tallied. We calculated inter-subject agreement for an image pair as the fraction of subjects who agreed with the majority decision, i.e.

$$Agreement = \frac{max(N_L, N_R)}{N} \quad (3)$$

where N_L and N_R are the number of subjects who chose the left and right objects, and $N = 35$ is the total number of subjects. This value is 1 when all subjects agree, and 0.514 when there is a 17-18 split. A histogram of the agreement values is shown in Figure 17. For each image pair, we also computed the majority image as the image selected by the majority of subjects. This allowed us to evaluate the accuracy of each person as the number of image pairs for which they selected the majority image. The average human accuracy for our experiment was 81.83%.

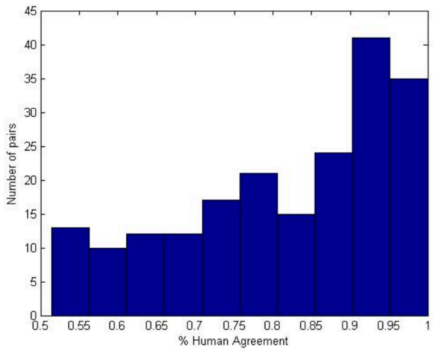


Figure 17: Agreement histogram over image pairs.

To determine the significance of the data, we looked at the number of subjects who voted "Left" on each of the 200 image pairs. An F-test revealed a statistically significant variance of these numbers, with a p-value of 5.31e-70. There was a significant negative effect of response time on agreement (p = 4.34e-11) but no effect of rectangular vs. convex polygonal objects (p = 0.564).

As mentioned in the previous section, we also evaluated the accuracy of several models in predicting the data. The results are shown in Table 1. The best of the SVM models (SVM3) was able to achieve a 74% accuracy. By contrast, the goal relevance model actually beat the human accuracy, with an 85% correct prediction rate.

Model	**Accuracy (%)**
SVM1	53.00
SVM2	68.50
SVM3	74.00
Human	81.83
Goal Relevance	85.00

Table 1: Prediction accuracy of models

To further examine the predictions, we divided the image pairs into groups based on the human agreement for the pairs. By definition, the human accuracy will be higher on image pairs that have more agreement. Will the models follow the same pattern? This is shown in Figure 18. For the three regression lines, R^2 error measurements and p-values were computed. For the human regression line, an R^2 of 0.999 and a p-value of 1.59e-5 were obtained. The goal relevance model had an R^2 of 0.935 and a p-value of 0.007. Finally, the SVM3 line had values of 0.074 and 0.0515, respectively, making it the only line that did not pass the significance threshold, although it was very close.

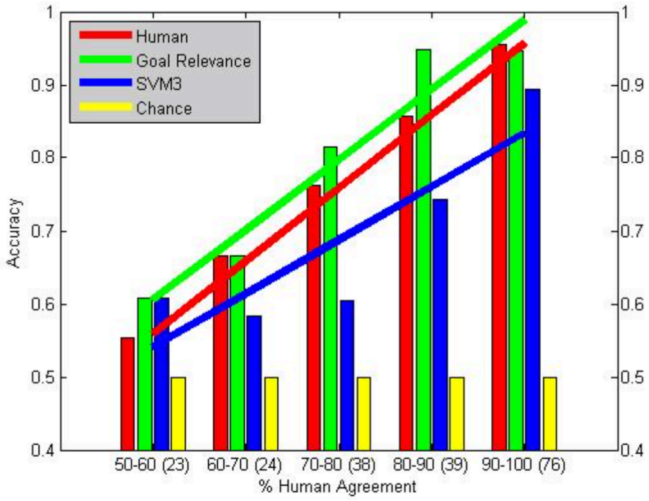


Figure 18: Model accuracies for image pairs with different levels of human agreement, and corresponding regression lines. The number of image pairs in each category is shown on the X axis with the agreement values.

We also investigated whether the magnitude of difference computed for each object by the goal relevance model was correlated with the level of human agreement for each image pair. A scatter plot of these values is shown in Figure 19, along with a regression line. This line was also statistically significant, with R^2 and p values of 0.148 and 1.885e-8, respectively.

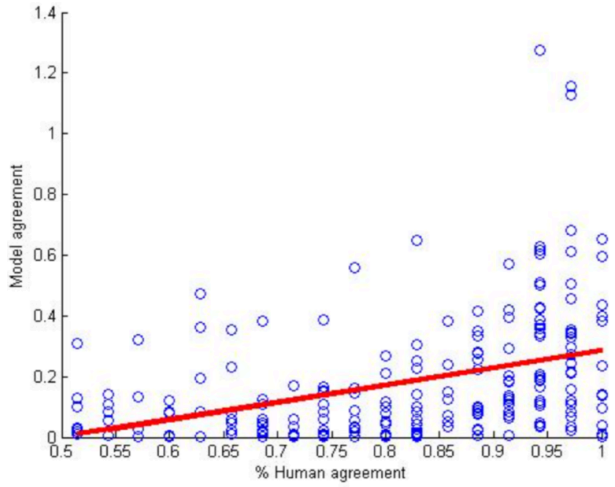


Figure 19: Scatter plot of goal relevance differences and human agreements for each image pair, and the regression line.

3.5. Conclusions on goal relevance measure

Our results show a clear advantage in using the proposed concept of goal relevance to predict human responses in this task. The model outperformed all of the SVM variants we constructed. Also, it found the same image pairs easy or challenging as did the humans, indicated by the similar slopes of the regression lines in Figure 18 and the significant slope of the line in Figure 19.

It is interesting that goal relevance was even able to outperform the average human accuracy. This may be because goal relevance, despite being an attempt to approximate the human notion of task-based relevance, might still be more precise in its' process and calculations. Humans may utilize heuristics or other methods for determining the available solutions to problems, and for detecting changes in that space. Also, participants may have tried to avoid spending a large amount of time on any image pair, eventually selecting a random answer which would decrease performance. However, our results suggest that the key idea behind goal relevance is sound; observations are relevant to a problem when they change the set of solutions. This is demonstrated by the success of the model.

There are a few potential issues with this study. Firstly, the theory does not define how goal relevance can be computed when the prior or posterior solution set is empty. In this case the KL divergence is not defined, so a natural interpretation using Eqn. (2) is not available. None of our image pairs contained this case, so it was not dealt with in our data. Another potential problem is that in constructing the SVM models, we could have missed a very informative feature. Lastly, an issue with the implementation is that the RRT path sampling method contains the well-known Voronoi bias. We attempted to minimize this by swapping the start and goal locations for half of the samples, which eliminated the non-symmetric aspect of the bias. However, it could still have an impact on the results.

We have described a theory of object relevance in the context of tasks, which has been termed goal relevance. Also, we have shown that goal relevance is successful in predicting human responses when asked to select task-relevant objects. To our knowledge, this is the first model to capture this concept in a quantitative manner. The results of the present study support the idea of information utility discussed in [7]. Further research will hopefully show that this theory generalizes to other types of tasks.

4. Computing surprise over web pages and social media feeds

We describe two completely new applications of our general mathematical theory of surprise, in domains and data types never explored before:

A. Finding surprising web pages
B. Finding fake followers (software robots as opposed to humans) in social media

4.A. Finding surprising web pages

As mentioned in the original proposal for this work, one of the thrusts was to try to apply our surprise and relevance theories to other large-scale data streams than the video streams investigated to date. Here we describe progress on finding surprising web pages. The general workflow is as follows:

- Collect several corpora of web pages (e.g., pages in English vs in French; real pages vs. web pages generated randomly from a dictionary of words).
- Parse the pages to strip out all HTML markup and focus on the core text information.
- Compute features over the text content, here using Latent Dirichlet Allocation (LDA) to discover latent (hidden variable) "topics" that underly each page.
- Compute surprise over the distributions of features obtained for the different corpora.

Data collection: To create corpora, we downloaded pages from the web (typically, the top 100 search results for a given query). We developed a web crawler tool in java to collect html pages from different search engines (Google, Bing, Baidu, etc). By specifying keywords and number of pages we want, the tool can automatically fetch the top search results and store them in local directories named by the corresponding search keywords.

As a control, we also generated pseudo-random English and French text using a Markov chains algorithm. Our method was essentially adopted from the book "Programming Pearls", section 15.3, Generating Text.

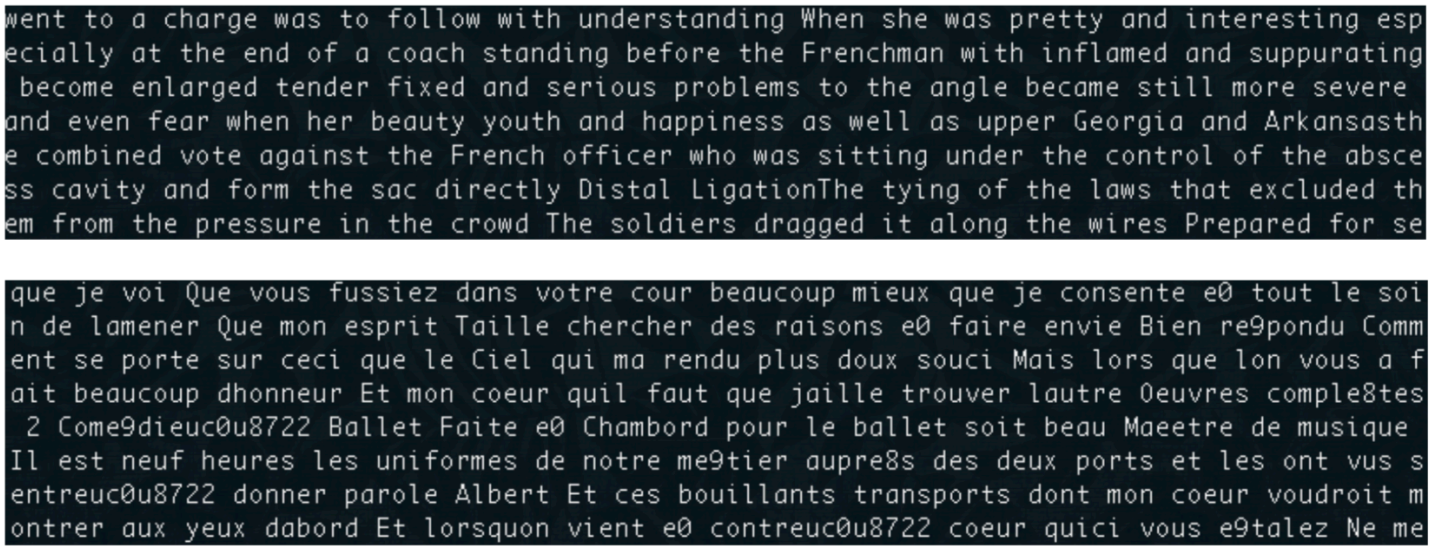


Figure 20: Random text generated using the Markov chain algorithm. (top) English. (bottom) French.

Parsing and stripping markup out: Before learning LDA by Gibbs Sampling, we need to perform the following preprocessing on the generated/collected raw data to focus on the text content only as opposed to markup and formatting:

• change words to lower case

• remove special characters

• remove English stop words

• stem words (do, doing, did, done are all replaced by do)

• remove URLs

• remove empty files

• remove javascript code

• remove CSS style markups

We developed some custom software to achieve this.

Computing features over each page: Surprise and relevance computation in our general mathematical frameworks operate by computing differences between prior and posterior distributions of beliefs in some feature space. Thus, an essential step in applying these theories to new data types is to find suitable features that can summarize and represent the raw data. In the web search community, many different features have been developed, which are used to rank-order search results. These start from the simplest (e.g., number of occurrences of a given word in a given page) to more complex machine-learning features (which often are proprietary). Here, we applied LDA (Latent Dirichlet Allocation) model to analyze all the files, random and meaningful, to gener-

ate topic models and learn the files' topic distribution. In the LDA framework, each document is modeled as representing a mixture of several different topics. To avoid having to define topics manually (which can be quite tedious, subjective, and arbitrary), the algorithm discovers topics, assigns probabilities or weights to the different topics for each document, and associates words in the documents with the underlying topics.

For surprise computation, we used theta, the topic matrix, as the likelihood distribution function.

The features used for surprise computation hence essentially look as follows:

	[,1]	[,2]	[,3]	[,4]	[,5]
1xaSKh-R.txt	0.016276704	0.010172940	0.000000000	0.02644964	0.03763988
1Y52wf-R.txt	0.001014199	0.001014199	0.000000000	0.03245436	0.04969574
1yBgpG-R.txt	0.000000000	0.000000000	0.003021148	0.05236657	0.02819738
1YjN36-R.txt	0.007135576	0.028542304	0.003058104	0.05402650	0.02854230
1yR6Ri-R.txt	0.000000000	0.003030303	0.001010101	0.07474747	0.03737374
1ySlZy-R.txt	0.015274949	0.012219959	0.000000000	0.05702648	0.03258656
1YZdtf-R.txt	0.004052685	0.016210740	0.005065856	0.03951368	0.02836879
1zoFBj-R.txt	0.001014199	0.013184584	0.000000000	0.07200811	0.03549696
1zt3ls-R.txt	0.003021148	0.007049345	0.007049345	0.04531722	0.05740181
1zyUJC-R.txt	0.036548223	0.003045685	0.000000000	0.07208122	0.03654822

	[,6]	[,7]	[,8]	[,9]	[,10]
1xaSKh-R.txt	0.017293998	0.6836216	0.03967447	0.13733469	0.031536114
1Y52wf-R.txt	0.006085193	0.6987830	0.05273834	0.11866126	0.039553753
1yBgpG-R.txt	0.046324270	0.6918429	0.05840886	0.08559919	0.034239678
1YjN36-R.txt	0.031600408	0.6687054	0.04689093	0.10499490	0.026503568
1yR6Ri-R.txt	0.017171717	0.7131313	0.03232323	0.09090909	0.030303030
1ySlZy-R.txt	0.020366599	0.7301426	0.05091650	0.08146640	0.000000000
1YZdtf-R.txt	0.014184397	0.7325228	0.04052685	0.09219858	0.027355623
1zoFBj-R.txt	0.003042596	0.6997972	0.02738337	0.12373225	0.024340771
1zt3ls-R.txt	0.011077543	0.6576032	0.04733132	0.15810675	0.006042296
1zyUJC-R.txt	0.001015228	0.7299492	0.02842640	0.07005076	0.022335025

Figure 21: Example where 10 topics were discovered (numbers in square brackets) to explain a corpus of documents. Each topic is associated with some underlying concept or collection of concepts, as discovered by the LDA algorithm. For each document (the xxx.txt files above), we then obtain a probability that a given topic contributed to this particular document (the numerical entries in the table).

Computing surprise: We computed the surprise value of each document based on the LDA topic matrix obtained in the previous step. Specifically, each row of the LDA matrix serves as the likelihood distribution function when we apply the Bayesian surprise framework.

The surprise value is define in our theory (see year 1 progress report) as the KL distance between posterior belief (here, the current page i) and prior belief (here, established from all the other pages in a corpus except i). Based on how prior and posterior belief is computed, we can compute surprise in the following two ways:

1) From average LDA to Surprise:

a. First compute the average topic distribution of all the documents except for the current page i, and use this average distribution to generate prior beliefs for page i;

b. compute i's posterior belief using the i th row in the LDA matrix;

c. compute surprise of page i by comparing prior (a) and posterior (b) belief (KL distance).

2) Pairwise document surprise followed by average:

Alternatively, we can also establish the prior for a given page by using only one other page (as opposed to all the others). We then compute surprise of file i by comparing i with each of the rest of the files iteratively and, in the end, computing the average value obtained for all the pairs.

Evaluation: At the time of this writing, thorough evaluation is still under way. Here we show some preliminary results for a preview:

Finding random pages among pages collected from the web (as search results): Here we mixed 25 randomly generated documents with 100 pages collected from the web. Total dataset size is hence 125 documents. On average over multiple runs (using different random and real pages) in our testing so far, 70% of the Top 10 sur-

prising files (7 out of 10) are random files, i.e., our algorithm was successful at pulling out random pages from real pages. In addition, 74% of Top 20 (15 out of 20) surprising files are random files, and 70% f the Top 30 surprising files (21 out of 30) are random files.

Compared to chance level, which is 25 / (100+25) = 0.2, the probability of finding a random page among a mix of real and random pages using our surprise theory is over 350% better than chance.

In testing so far, finding French pages among a mixture of English and French pages is too easy (all of the top 25 surprising pages are the ones in French, when we mix 25 French pages with 100 English pages). Thus we are now moving to more difficult testing where all pages will be in one language, but either collected using slightly different keyword searches, or generated using slightly different word probabilities in the Markov chain model.

4.B. Finding fake followers (software robots as opposed to humans) in social media

The work aims at distinguishing real users from fake users (created by automated computer programs) in twitter-like social platforms (for testing, we used Weibo). Fake followers are accounts created to inflate the number of followers of a target account. Buying fake followers is actually a quite common commercial behavior. Many "celebrities" are followed by thousands or millions of fans, and significant profit will be gained if these "celebrities" post an advertising message or other event promotion. Many people who initially have only a small number of followers or fans intend to follow this pattern, i.e., they acquire multitudes of fake followers on the Internet. Making them more trustworthy and influential, they can easily attract more genuine followers and profits will follow. This phenomenon is common in both twitter and weibo. These platforms also perform large-scale fake fans deletion from time to time. As a result, the fake followers are produced with more care. They are hard to single out by the naked eye from their photo, nick name, or even message content.

Figure 22: Two profiles on weibo (Chinese twitter). The one on the left is a real person while the one on the right is fake (created by a software agent).

Recently the detection approaches for fake followers focus on the relationship network and some other traces. Here we have an idea to detect them simply by the content. The normal people will publish topics they are interested in. Fake followers' content is a random copy of others', thus shows a huge variance in topic distribution. For example, the abnormal content may be copied from a mom and a student, covering both the school life and child parenting.

Assumptions:

◆ Normal people only focus on some specific topics
◆ Fake followers simply copy content from normal people at random

Example:

Suppose there are 5 probable topics in total. [School, Child, Sales, Travel, Food]

The topic distribution of a fake follower may be [0.2, 0.2, 0.2, 0.2, 0.2] because it copied content from different people with uniformly distributed probability (i.e., covers the 5 topics each with a probability 1/5).

In contrast, the distribution of normal people, for instance, a student, may be [0.5, 0, 0.2, 0.2, 0.1] because s/he pays more attention to school life but does not yet have a child to talk about.

Simple Instance of Topic Distribution

	School	Child	Sales	Travel	Food
Student	0.5	0	0.1	0.2	0.2
Mom	0.1	0.6	0.2	0.1	0.1
Fake Followers	0.2	0.2	0.2	0.2	0.2

Dataset:
- 500 fake followers were purchased on the Internet.
- 1000 normal people are from a social network prediction competition.
- Each file is a composition of all the weibo content of one person. All the files are filtered to guarantee that each one contains 150~200 messages, around 10 thousand words.

Pre-processing:

1. Extract only nouns. We only extract noun words from the corpus because noun words are easier to pair with topics. We use the github/jieba library for tokenizing.
2. Translate into English. Although the dataset is composed of Chinese weibo, we also want to prove feasibility on Twitter. So all the files were translated into English via translation api.
3. Delete meaningless words. We deleted all the meaningless word like “the” ,”of” etc and redundant spaces.

Apply LDA to the dataset:
This part is similar to that in the project above (section A). The number of general topic in weibo content is 15-20, so we set K (the number of topics to calculate LDA) to 15.

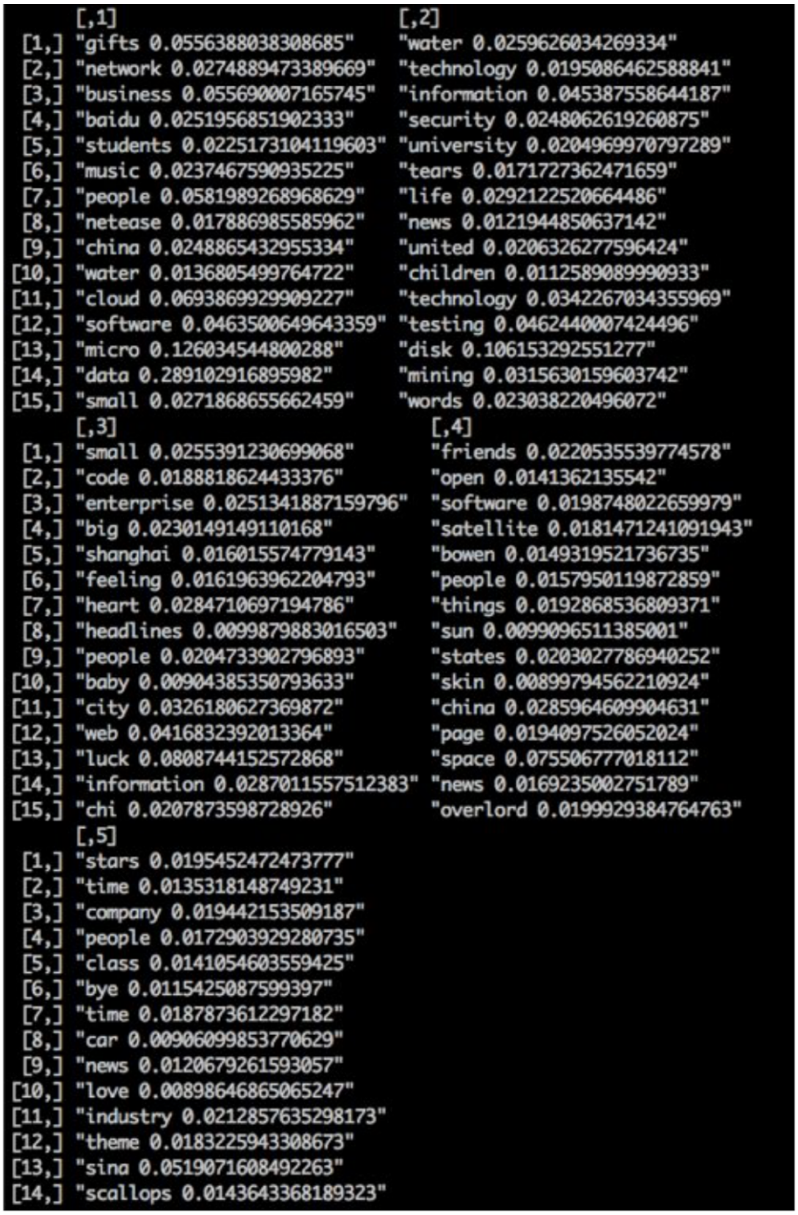


Figure 23: The top 5 words for each of the 15 topics discovered by LDA.

Calculating Surprise: The approach is similar to that of Section A of this report:

a. Prior belief of Page K: average topic distribution of all the files except for Page K.
b. Compute K's posterior belief using the ith row in the LDA matrix.
c. Compute surprise of Page K by comparing prior (a) and posterior (b) belief.
d. Sort all the files in ascending order of surprise value.

Results:

Interestingly, in our surprise ranking, the top 17 files (most surprising) belonged to the "normal people" set. We thus checked these files one by one, only to find that they are actually fake followers. Hence, he dataset from the official competition failed to filter them out as fake follower while our approach found them.

The top 500 most surprising people are almost all from fake followers, the last 1000 are almost all from normal people. Since the total number of fake followers' files is 500, and they actually take up 450 files in the top 500 rank-ordered by surprise, the accuracy rate is 450/500*100=90%.

5. Applications and basic science progress

In addition to core basic research activities related to defining goal relevance, we have explored several application domains where attention, surprise, and goal relevance are important. Brief highlights from these studies are presented below.

We have continued to push further our understanding of human vision, to enable the next generation of vision-based algorithms for driver assistance. This year we have three main results:

A new gaze prediction model that learns the relationship between saliency maps and fixations from examples (Zhao et al., Proc IEEE CVPR, 2015): Predicting where humans will fixate in a scene has many practical applications. Biologically-inspired saliency models decompose visual stimuli into feature maps across multiple scales, and then integrate different feature channels, e.g., in a linear, MAX, or MAP. However, to date there is no universally accepted feature integration mechanism. Here, we propose a new a data-driven solution: We first build a "fixation bank" by mining training samples, which maintains the association between local patterns of activation, in 4 feature channels (color, intensity, orientation, motion) around a given location, and corresponding human fixation density at that location. During testing, we decompose feature maps into blobs, extract local activation patterns around each blob, match those patterns against the fixation bank by group lasso, and determine weights of blobs based on reconstruction errors. Our final saliency map is the weighted sum of all blobs. Our system thus incorporates some amount of spatial and featural context information into the location-dependent weighting mechanism.

Enhancing bottom-up visual saliency models with higher-level features (Borji et al., J Vis, 2014; Parks et al., Vis Res 2014): Saliency models typically analyze images along low-level feature dimensions (e.g., color contrast, oriented edges) to make predictions of what might attract the gaze of a driver. Here we investigated how higher-level features could be integrated to such models. The high-level feature we considered is gaze following, the tendency people have to follow the gaze of actors in scenes, to look towards what those actors are looking at. Indeed, previous studies have shown that gaze direction of actors in a scene influences eye movements of passive observers during free-viewing. However, no computational model has been proposed to combine bottom-up saliency with actor's head pose and gaze direction for predicting where observers look. Here, we first learn probability maps that predict fixations leaving head regions (gaze following fixations), as well as fixations on head regions (head fixations), both dependent on the actor's head size and pose angle. We then learn a combination of gaze following, head region, and bottom-up saliency maps with a Markov chain composed of head region and non-head region states. This simple structure allows us to inspect the model and make comments about the nature of eye movements originating from heads as opposed to other regions. Here, we assume perfect knowledge of actor head pose direction (from an oracle). The combined model, which we call the Dynamic Weighting of Cues model (DWOC), explains observers' fixations significantly better than each of the constituent components. Finally, in a fully automatic combined model, we replace the oracle head pose direction data with detections from a computer vision model of head pose. Using these (imperfect) automated detections, we again find that the combined model significantly outperforms its individual components. Our work extends the engineering and scientific applications of saliency models and helps better understand mechanisms of visual attention.

Teaching and outreach activities: We taught a course at SIGGRAPH 2014 (#1 computer graphics conference) about attention modeling and applications to computer graphics (McNamara et al., 2014).

Applications to analysis of eye movements (Borji & Itti, J. Vision, 2014): In a very influential yet anecdotal illustration, Yarbus suggested that human eye movement patterns are modulated top-down by different task demands. While the hypothesis that it is possible to decode the observer's task from eye movements has received some support (e.g., Iqbal & Bailey (2004); Henderson et al. (2013)), Greene et al. (2012) argued against it by reporting a failure. In this study, we perform a more systematic investigation of this problem, probing a larger number of experimental factors than previously. Our main goal is to determine the informativeness of eye movements for task and mental state decoding. We perform two experiments. In the first experiment, we re-

analyze the data from a previous study by Greene et al. (2012) and contrary to their conclusion, we report that it is possible to decode the observer's task from aggregate eye movement features slightly but significantly above chance, using a Boosting classifier (34.12% correct vs. 25% chance-level; binomial test, p=1.0722e-04). In the second experiment, we repeat and extend Yarbus' original experiment by collecting eye movements of 21 observers viewing 15 natural scenes (including Yarbus' scene) under Yarbus' seven questions. We show that task decoding is possible, also moderately but significantly above chance (24.21% vs. 14.29% chance-level; binomial test, p=2.4535e-06). We thus conclude that Yarbus' idea is supported by our data and continues to be an inspiration for future computational and experimental eye movement research. From a broader perspective, we discuss techniques, features, limitations, societal and technological impacts, and future directions in task decoding from eye movements.

Applications to robotics navigation (Siagian et al., J Field Robotics, 2014): We have continued to investigate how the concepts of visual attention, surprise, relevance, visual salience, bio-inspired object recognition, and rapid computation of the "gist" of a scene can provide better environmental awareness to autonomous and semi-autonomous vehicles, so as to enable better goal-oriented behavior.

This is the first time a robot that relies primarily on neuro-inspired visual processing is demonstrated to perform so well during such large-scale, real-world testing. Many more details are available in (Siagian, Chang & Itti, J Field Robotics, 2014). We are planning to more tightly integrate our new definition of relevance to the core algorithms running on this robot.

Applications to non-visual data streams (Windau & Itti, IROS 2013): Beyond applications of our theory and algorithm work to video data streams, we also explored how simple data streams – collected from accelerometers mounted on glasses – could be used to decode user activities and intentions. This requires that the most surprising and relevant events be first mined from within the constant stream of sensor data. With the advent of wearable computing devices like Google Glass, which integrate video camera, small display, and accelerometers to a glasses frame, this type of user monitoring may become widely available in the near future. In fact, after successful demonstration of a prototype described below, we received a small research grant from Google to implement our algorithms on Google Glass.

A paper describing our results was presented at IROS 2013 (Windau & Itti, 2013).

References

[1] Ballard, Dana H., and Mary M. Hayhoe. ”Modelling the role of task in the control of gaze.” Visual Cognition 17.6-7 (2009): 1185-1204.

[2] Borji, Ali and Sihite, Dicky Nauli. and Itti, Laurent, What/Where to Look Next? Modeling Top-down Visual Attention in Complex Interactive Environments, IEEE Transactions on Systems, Man, and Cybernetics, Part A - Systems and Humans (2012):44(5):523-538

[3] Buswell, Guy Thomas. ”How people look at pictures: a study of the psychology and perception in art.” (1935).

[4] Castelhano, Monica S., Michael L. Mack, and John M. Henderson. ”Viewing task influences eye movement control during active scene perception.” Journal of Vision 9.3 (2009).

[5] Churchill, Alexander L. et al. ”Twitter Relevance Filtering via Joint Bayes Classifiers from User Clustering.” (2010).

[6] Gigerenzer, Gerd. ”Why heuristics work.” Perspectives on Psychological Science 3.1 (2008): 20-29.

[7] Gottlieb, Jacqueline, and Puiu Balan. ”Attention as a decision in information space.” Trends in cognitive sciences 14.6 (2010): 240-248.

[8] Hayhoe, Mary, and Dana Ballard. ”Eye movements in natural behavior.” Trends in cognitive sciences 9.4 (2005): 188-194.

[9] Hj_rland, Birger. ”The foundation of the concept of relevance.” Journal of the American Society for Information Science and Technology 61.2 (2010): 217-237.

[10] Hj_rland, Birger, and Frank Sejer Christensen. ”Work tasks and socio-cognitive relevance: A specific example.” Journal of the American Society for Information Science and Technology 53.11 (2002): 960-965.

[11] Huang, Liqiang, and Harold Pashler. ”Quantifying object salience by equating distractor effects.” Vision research 45.14 (2005): 1909-1920.

[12] Itti, Laurent, and Pierre Baldi. ”Bayesian surprise attracts human attention.” Advances in neural information processing systems 18 (2006): 547.

[13] LaValle, Steven M. ”Rapidly-Exploring Random Trees: A New Tool for Path Planning.” (1998).

[14] Malcolm, George L., and John M. Henderson. ”Combining top-down processes to guide eye movements during real-world scene search.” Journal of Vision 10.2 (2010): 4.

[15] Navalpakkam, Vidhya, and Laurent Itti. ”Modeling the influence of task on attention.” Vision research 45.2 (2005): 205-231.

[16] Peters, Robert J., and Laurent Itti. ”Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention.” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[17] Rothkopf, Constantin A., Dana H. Ballard, and Mary M. Hayhoe. ”Task and context determine where you look.” Journal of Vision 7.14 (2007): 16.